

QUANTIZATION TRAINING, IMAGE PROCESSING METHOD AND APPARATUS, STORAGE MEDIUM

CROSS-REFERENCE TO RELATED APPLICATION

[0001] This application claims priority to Chinese Patent Application No. 202010434807.1, filed on May 21, 2020 and entitled "QUANTIZATION TRAINING, IMAGE PROCESSING METHOD AND APPARATUS, STORAGE MEDIUM", the disclosure of which is hereby incorporated by reference in its entirety.

TECHNICAL FIELD

[0002] The present disclosure relates to the field of quantization training, and in particular to quantization training, an image processing method and apparatus, and a storage medium.

BACKGROUND

[0003] As more and more neural network models need to be deployed on mobile devices, the efficiency of inference has become a key issue. When a neural network model is deployed to a mobile device, the structure of the neural network model needs to be simplified, and the usual methods include quantization.

[0004] Quantization of the neural network model is to approximate the high-precision parameters of the neural network model with lower-precision parameters. The high-precision parameters may include floating-point parameters, and the low-precision parameters may include integer parameters. The quantified neural network model may process more data per unit time, and further reduce the storage space occupied, and so on.

[0005] At present, the quantization training process on neural network models generally specifies the total number of quantization training according to experience. After the total number of times is reached, the obtained quantization model is converted into the corresponding test model in the actual hardware environment, and the test model is run in the actual hardware environment to obtain the running result.

SUMMARY

[0006] According to a first aspect in the embodiments of the present disclosure, a quantization training method is provided. The method includes the following operations. A model training device performs at least one round of iterative quantization training on a neural network model to obtain a first quantization model. The model training device tests the first quantization model by simulating a hardware deployment environment to obtain a test result of the first quantization model.

[0007] In some optional embodiments, the operation that the model training device tests the first quantization model by simulating the hardware deployment environment to obtain the test result of the first quantization model may include the following action. The model training device tests the first quantization model by invoking a target function to obtain the test result of the first quantization model. The target function is used to simulate the hardware deployment environment.

[0008] In some optional embodiments, the method may further include the following operation. Conversion processing is performed on the first quantization model to obtain a first test model. The conversion processing includes removing at least one target unit of the first quantization model, and the target unit is configured to perform quantization and/or dequantization on at least one of output data or network parameters of a network layer of the neural network model. The operation that the model training device tests the first quantization model by simulating the hardware deployment environment to obtain the test result of the first quantization model may include the following action. The model training device tests the first test model by simulating the hardware deployment environment to obtain the test result of the first quantization model.

[0009] In some optional embodiments, the conversion processing may be implemented by a target function used for simulating the hardware deployment environment.

[0010] In some optional embodiments, the operation that the model training device tests the first quantization model by simulating the hardware deployment environment to obtain the test result of the first quantization model may include the following action. The model training device tests the first quantization model by simulating the hardware deployment environment and using fixed-point data obtained by fixed-point processing on test samples and network parameters of the first quantization model to obtain the test result of the first quantization model.

[0011] In some optional embodiments, the method may further include the following operations. The model training device performs at least one round of iterative quantization training on the first quantization model to obtain a second quantization model. The model training device tests the second quantization model by simulating the hardware deployment environment to obtain a test result of the second quantization model.

[0012] In some optional embodiments, the method may further include the following operation. A training strategy analysis result for quantization training of the neural network model is obtained based at least in part on the test result of the first quantization model. The training strategy analysis result includes at least one of the following: terminating the quantization training on the neural network model, adjusting a quantization manner of at least one network layer in the neural network model, or adjusting a manner of the subsequent iterative quantization training of the neural network model.

[0013] In some optional embodiments, the operation that the model training device performs at least one round of iterative quantization training on the first quantization model to obtain the second quantization model may include the following action. The model training device performs at least one round of iterative quantization training on the first quantization model in parallel with the process of testing the first quantization model, to obtain the second quantization model.

[0014] In some optional embodiments, the operation of testing the first quantization model to obtain the test result of the first quantization model may include any one of the following: the obtained first quantization model is tested responsive to that at least one round of iterative quantization training on the neural network model reaches a preset number of times to obtain the test result of the first quantization model; or, the first quantization model is tested responsive to determining that the first quantization model satisfies a test condition based on a preset test strategy to obtain the test result of the first quantization model.

[0015] According to a second aspect in the embodiments of the present disclosure, an image processing method is provided. The method includes the following operation. An image to be processed is input into a quantization model to obtain an image processing result output by the quantization model. The first quantization model is a quantization model obtained through the method according to the first aspect.

[0016] According to a third aspect in the embodiments of the present disclosure, a quantization training apparatus is provided. The apparatus includes a first quantization training module and a first test module. The first quantization training module is configured to perform at least one round of iterative quantization training on a neural network model by a model training device to obtain a first quantization model. The first test module is configured to test the first quantization model through simulating a hardware deployment environment by the model training device to obtain a test result of the first quantization model.

[0017] In some optional embodiments, the first test module may include a first test submodule. The first test submodule is configured to test the first quantization model through invoking a target function by the model training device to obtain the test result of the first quantization model. The target function is used to simulate the hardware deployment environment.

[0018] In some optional embodiments, the apparatus may further include a model conversion module. The model conversion module is configured to perform conversion processing on the first quantization model to obtain a first test model. The conversion processing includes removing at least one target unit of the first quantization model, and

the target unit is configured to perform quantization and/or dequantization on at least one of output data or network parameters of a network layer of the neural network model. The first test module may include a second test submodule. The second test submodule is configured to test the first test model through simulating the hardware deployment environment by the model training device to obtain the test result of the first quantization model.

[0019] In some optional embodiments, the conversion processing may be implemented by a target function used for simulating the hardware deployment environment.

[0020] In some optional embodiments, the first test module may include a third test submodule. The third test submodule is configured to test the first quantization model through simulating the hardware deployment environment and using fixed-point data obtained by fixed-point processing on test samples and network parameters of the first quantization model by the model training device to obtain the test result of the first quantization model.

[0021] In some optional embodiments, the apparatus may further include a second quantization training module and a second test module. The second quantization training module is configured to perform at least one round of iterative quantization training on the first quantization model by the model training device to obtain a second quantization model. The second test module is configured to test the second quantization model through simulating the hardware deployment environment by the model training device to obtain a test result of the second quantization model.

[0022] In some optional embodiments, the apparatus may further include a determination module. The determination module is configured to obtain a training strategy analysis result for quantization training of the neural network model based at least in part on the test result of the first quantization model. The training strategy analysis result includes at least one of the following: terminating the quantization training on the neural network model, adjusting a quantization manner of at least one network layer in the neural network model, or adjusting a manner of the subsequent iterative quantization training of the neural network model.

- [0023] In some optional embodiments, the second quantization training module may include a quantization training submodule. The quantization training submodule is configured to perform at least one round of iterative quantization training on the first quantization model in parallel with the process of testing the first quantization model by the model training device, to obtain the second quantization model.
- [0024] In some optional embodiments, the first test module may include any one of a fourth test submodule and a fifth test submodule. The fourth test submodule is configured to test the obtained first quantization model responsive to that at least one round of iterative quantization training on the neural network model reaches a preset number of times to obtain the test result of the first quantization model. The fifth test submodule is configured to test the first quantization model responsive to determining that the first quantization model satisfies the test condition based on a preset test strategy to obtain the test result of the first quantization model.
- [0025] According to a fourth aspect in the embodiments of the present disclosure, an image processing apparatus is provided. The apparatus includes an image processing module, configured to input an image to be processed into a quantization model to obtain an image processing result output by the quantization model. The quantization model is a quantization model obtained through the method according to the first aspect.
- [0026] According to a fifth aspect in the embodiments of the present disclosure, a computer-readable storage medium is provided. The storage medium stores a computer program, and the computer program is used to execute the quantization training method according to the first aspect or the image processing method according to the second aspect.
- [0027] According to a sixth aspect in the embodiments of the present disclosure, a quantization training apparatus is provided. The apparatus includes a processor, and a memory for storing an executable instruction for the processor. The processor is configured to invoke the executable instruction stored in the memory to implement the quantization training method according to the first aspect.
- [0028] According to a seventh aspect in the embodiments of the present disclosure, an

image processing apparatus is provided. The apparatus includes a processor, and a memory for storing an executable instruction for the processor. The processor is configured to invoke the executable instruction stored in the memory to implement the image processing method according to the second aspect.

[0029] The technical solutions provided by the embodiments of the present disclosure may include the following beneficial effects.

[0030] In the embodiments of the present disclosure, the model training device performs at least one round of iterative quantization training on a neural network model to obtain a first quantization model, and further, the model training device tests the first quantization model by simulating a hardware deployment environment to obtain a test result of the first quantization model. In the present disclosure, the model training device may simulate the hardware deployment environment after at least one round of iterative quantization training is performed, and directly test the first quantization model under the training framework,. In such case, the test result of the first quantization model may be obtained without deploying the first quantization model into the actual hardware device, which is beneficial to improve the development efficiency of the neural network model.

[0031] It is to be understood that the above general descriptions and detailed descriptions below are only exemplary and explanatory and not intended to limit the disclosure.

BRIEF DESCRIPTION OF THE DRAWINGS

[0032] The accompanying drawings, which are incorporated in and constitute a part of this specification, illustrate embodiments consistent with the present disclosure and, together with the description, serve to explain the principles of the present disclosure.

[0033] FIG. 1 illustrates a flowchart of a quantization training method according to an exemplary embodiment of the present disclosure.

[0034] FIG. 2 illustrates a flowchart of another quantization training method

according to an exemplary embodiment of the present disclosure.

[0035] FIG. 3A illustrates an architecture diagram of a neural network model according to an exemplary embodiment of the present disclosure.

[0036] FIG. 3B illustrates an architecture diagram of a first quantization model according to an exemplary embodiment of the present disclosure.

[0037] FIG. 3C illustrates an architecture diagram of a first test model according to an exemplary embodiment of the present disclosure.

[0038] FIG. 4 illustrates a flowchart of another quantization training method according to an exemplary embodiment of the present disclosure.

[0039] FIG. 5 illustrates a flowchart of another quantization training method according to an exemplary embodiment of the present disclosure.

[0040] FIG. 6 illustrates a block diagram of a quantization training apparatus according to an exemplary embodiment of the present disclosure.

[0041] FIG. 7 illustrates a schematic structure diagram of a quantization training apparatus according to an exemplary embodiment of the present disclosure.

DETAILED DESCRIPTION

[0042] Exemplary embodiments will be described in detail, examples of which are illustrated in the accompanying drawings. When the following description refers to the drawings, unless otherwise indicated, the same numbers in different drawings indicate the same or similar elements. The implementations described in the following exemplary embodiments do not represent all implementations consistent with the present disclosure. Instead, they are merely examples of apparatuses and methods consistent with some aspects of the present disclosure as detailed in the appended claims.

[0043] The terms used in the present disclosure are only for describing specific embodiments, and are not intended to limit the present disclosure. The singular forms of

"a", "the" and "this" used in the present disclosure and appended claims are also intended to include plural forms, unless the context clearly indicates other meanings. It should also be understood that the term "and/or" used herein refers to and includes any or all possible combinations of one or more associated listed items.

[0044] It should be understood that although the terms first, second, third, etc. may be used in the present disclosure to describe various information, but the information should not be limited to these terms. These terms are only used to distinguish the same type of information from each other. For example, without departing from the scope of the present disclosure, the first information may also be referred to as second information, and similarly, the second information may also be referred to as first information. Depending on the context, the word "if" used herein can be interpreted as "when" or "while" or "in response to determining".

[0045] At present, the neural network model needs to be deployed on actual mobile devices for testing after the quantization training is completed. Specifically, before the test, the neural network model may be performed an iterative quantization training, after the quantization model is obtained, the quantization model is converted into a test model corresponding to the actual hardware environment, and then the test model is deployed to the actual mobile device for test and the test result is obtained. According to the test result, it is determined whether the performance of the quantization model meets expectations.

[0046] The quantization scheme may generally be divided into Post-training quantization and Quantization-aware training. Post-training quantization refers to direct quantization on parameters of the neural network model after training the neural network model with floating-point parameters. This scheme has better quantization effects on models with a large number of parameters and less performance loss, but lead to significant performance degradation for models with a small number of parameters. Quantization-aware training is to simulate a quantization behavior in the training of the neural network model, store the fixed-point parameters in the form of floating-point parameters in the training and directly use the fixed-point parameters for calculation during the final model inference.

- [0047] The process of quantization training is introduced by taking Quantization-aware training as an example. Firstly, all tensor operations are implemented in modules, and then all reused modules are changed to individual modules, that is, the module reuse is not allowed, and all functional interfaces in the training framework (such as pytorch) are changed to modules that implement the same function.
- [0048] However, the quantization model needs to be obtained firstly through a complete training process, then the trained quantization model is converted into a deployable model and placed in the deployment environment to test its performance in the deployment environment. This scheme is extremely inconvenient for model training or parameter adjustment, and needs to ensure that there is no error in the intermediate conversion. In addition, the model needs to be fully trained before the result of the model in the real environment may be known, and it is unable to correct the model training process as early as possible.
- [0049] In addition, the disadvantage of using training frameworks such as pytorch is that a quantization model including floating-point parameters is used for testing, while a quantization model including fixed-point parameters is used for inference in the deployment environment. Therefore, the test result may differ from the inference result on the deployment hardware, and the test result is unable to truly reflect the performance of the quantization model in the actual hardware deployment environment. In addition, the quantization model is prone to over-fitting during the training process, and once over-fitting occurs, it is difficult to adjust the parameters in the quantization model.
- [0050] Embodiments of the present disclosure provide a quantization training scheme. Such as illustrated in FIG. 1, FIG. 1 illustrates a quantization training method according to an exemplary embodiment, and the quantization training method may be applied to a model training device. In the embodiments of the present disclosure, both the quantization training and the testing of the quantization model may be performed on the same model training device. The model training device may use an electronic device deployed with a training platform framework including quantization algorithms or quantization tools. The electronic device may include, but is not limited to, a terminal device using x86 architecture, such as a personal computer (PC), a mobile phone, a

portable device, etc. The quantization training method includes the following operations.

[0051] In S101, the model training device performs at least one round of iterative quantization training on a neural network model to obtain a first quantization model.

[0052] The model training device may use quantization training methods, including but not limited to quantization-aware training, to quantify the neural network model in the training process. In the embodiments of the present disclosure, the model training device may perform at least one round of iterative quantization training on the neural network model to obtain the first quantization model. Specifically, the model training device may perform limited iteration training such as one or several times to obtain the first quantization model.

[0053] In S102, the model training device tests the first quantization model by simulating a hardware deployment environment to obtain a test result of the first quantization model.

[0054] In the embodiments of the present disclosure, the actual hardware deployment environment may be simulated on the model training device, so that the first quantization model may be tested directly on the model training device without converting the first quantization model to the actual hardware deployment environment for test. According to the test result of the first quantization model, the training strategy analysis of quantization training may be performed, for example, at least one round of iterative quantization training may be performed on the first quantization model again.

[0055] In some embodiments, the simulated hardware deployment environment refers to that at least one interface is encapsulated in a modular manner based on the operation logic of the hardware deployment environment. During testing, an interface corresponding to the network structure of the first quantization model in the at least one interface may be invoked to run the first quantization model to obtain the running result, and the running result is taken as the test result of the first quantization model. At least one interface is respectively used to implement the functions of different network layers, and the network layers include, but are not limited to, a convolution layer (conv), a pooling layer (pooling), a linear layer (linear), an activation function layer (prelu), etc.

[0056] For example, the first quantization model needs to run on a graphics processing unit (GPU) of a mobile device, and the hardware deployment environment is the hardware environment of the GPU. The functions of different network layers may be realized through at least one interface based on the calculation logic of the GPU. The model training device may invoke the interface corresponding to the network structure of the first quantization model in the at least one interface, and run the first quantization model to obtain the measurement result of the first quantization model.

[0057] In the above embodiment, the model training device may simulate the hardware deployment environment after performing at least one round of iterative quantization training, and directly test the first quantization model under the training framework. In this way, the test result of the first quantization model may be obtained without deploying the first quantization model on the actual hardware device, which is beneficial to improve the development efficiency of the neural network model.

[0058] In some optional embodiments, with respect to the operation in S102, the model training device may set at least one interface encapsulated in a modular manner based on the operation logic of the hardware deployment environment in the target function, that is, the target function may be used to simulate the hardware deployment environment. After the first quantization model is obtained, the interface corresponding to the network structure of the first quantization model in the target function may be invoked by means of function invoking, so as to obtain the test result of the first quantization model.

[0059] In a possible implementation, the quantization training framework of the model training platform adopts the pytorch framework, and correspondingly, the target function may be the function `model.quant()` in the quantization training framework.

[0060] In the above embodiments, the model training platform may invoke a target module to test the first quantization model and obtain the test result of the first quantization model. In this way, simulation of the hardware deployment environment on the same hardware device is realized, which may not only perform the iterative quantization training, but also directly obtain the test result of the first quantization model, thereby improving the development efficiency of the neural network model.

[0061] In some optional embodiments, during the iterative quantization training process, a pseudo quantization layer is added to the output of each network layer of the neural network model, and precision loss caused by the quantization of the network layer may be simulated by the pseudo quantization layer.

[0062] For example, original output data of the network layer 1 has the first precision, assumed that the first precision is FP32 with a value of 1.1, and the pseudo quantization layer quantizes the original output data to obtain the value of the second precision, assumed that the second precision is uint8 with a value of 1, but the network layer 2 that takes the output of the network layer 1 as input also needs to take the value of the first precision, and therefore, the value obtained by quantization of the network layer 1 needs to be dequantized by the pseudo quantization layer, so that the value 1 of the second precision is converted into the first precision, and the value remains unchanged. According to the above process, the precision loss of the network layer 1 may be determined by the pseudo-quantization layer to be $1.1-1=0.1$. In the same way, precision loss of other network layers may be determined.

[0063] In some optional embodiments, such as illustrated in FIG. 2, before the operation in S102 is performed, the method may further include the operation in S103.

[0064] In S103, conversion processing is performed on the first quantization model to obtain a first test model.

[0065] In the embodiment of the present disclosure, the pseudo quantization layer may be removed in the process of testing the first quantization model, and the same model structure as that in the actual deployment environment may be used for test, so that running of the neural network model obtained by quantization training on the actual deployment hardware may be directly simulated, and more realistic test result is obtained. Therefore, in the embodiments of the present disclosure, the first quantization model needs to be converted, the conversion processing includes removing at least one pseudo quantization layer of the first quantization model. In the embodiments of the present disclosure, the pseudo quantization layers may be implemented in a modular manner, that is, each pseudo quantization layer may correspond to a target unit, and correspondingly, the conversion processing includes removing at least one target unit of

the first quantization model, and the target unit is configured to perform quantization and/or dequantization on at least one of output data or network parameters of a network layer of the neural network model.

[0066] Correspondingly, the operation in S102 may include the following action.

[0067] The model training device tests the first test model by simulating the hardware deployment environment to obtain the test result of the first quantization model.

[0068] For example, the original network structure of the neural network model is illustrated in FIG. 3A, and the first quantization model including at least one target unit is illustrated in FIG. 3B. The target unit 1 performs quantization and/or dequantization on the network parameter of the convolution layer, that is, a weight value, and the target unit 2 performs quantization and/or dequantization on the output data of the activation function layer. The conversion processing needs to remove the target units from the first quantization model, and the first test model is obtained as illustrated in FIG. 3C.

[0069] In the present disclosure, the first test model is tested by the model training device, so as to obtain the test result of the first quantization model.

[0070] In the above embodiments, the first quantization model needs to perform quantization and/or dequantization on at least one of the output data or network parameters of the network layer in the neural network model. However, during the testing process, it is no longer necessary to perform quantization and/or dequantization operations. Therefore, the first quantization model may be converted to remove at least one target unit in the first quantization model to obtain the first test model, and test is performed with respect to the first test model, such that the required test result of the first quantization model may be obtained, thereby achieving testing of the first quantization model on the model training device.

[0071] In some optional embodiments, the process of converting the first quantization model by the model training device is implemented by a target function used to simulate the hardware deployment environment.

[0072] After obtaining the first quantization model, the model training device invokes

the target function to perform conversion processing on the first quantization model to remove at least one target unit, so as to test the obtained first test model to obtain the test result of the first quantization model.

[0073] In some optional embodiments, when the first quantization model is tested, fixed-point processing may be performed on test samples and network parameters of the first quantization model firstly to obtain fixed-point data, and then the first quantization model is tested by using the obtained fixed-point data to obtain the test result.

[0074] Fixed-point processing refers to conversion of data from the first precision to the second precision. The first precision is higher than the second precision, for example, the first precision is floating-point precision FP32, and the second precision is integer precision uint8.

[0075] The model training device needs to perform fixed-point processing on the test samples and network parameters, such as weight values, input from the first quantization model, and then obtains fixed-point data, and performs tests based on the fixed-point data.

[0076] For example, in FIG. 3B, the precision of the test sample (input value) and network parameter (a weight value corresponding to the input value) of the first quantization model are both FP32, and the precision of the fixed-point data after the fixed-point processing is uint8. The first test model in FIG. 3C may be tested based on the fixed-point data of uint8 to obtain the test result of the first quantization model.

[0077] In the above embodiments, the model training device needs to use the fixed-point data obtained by fixed-point processing on the test samples and network parameters of the first quantization model to test the first quantization model to obtain the test result of the first quantization model. In this way, the test process is more reasonable and accurate.

[0078] In some optional embodiments, such as illustrated in FIG. 4, the above method may further include the operations in S104 and S105.

[0079] In S104, the model training device performs at least one round of iterative

quantization training on the first quantization model to obtain a second quantization model.

[0080] In the embodiments of the present disclosure, after obtaining the first quantization model, the model training device may continue to perform at least one round of iterative quantization training on the first quantization model to obtain the second quantization model.

[0081] In a possible implementation, when the test result of the first quantization model indicates that the first quantization model does not meet the quantization training requirements, for example, when the accuracy of processing the test sample does not meet the design requirements on the neural network model, the model training device continues to perform at least one round of iterative quantization training on the first quantization model to obtain the second quantization model.

[0082] In another possible implementation, the model training device may also perform at least one round of iterative quantization training on the first quantization model to obtain the second quantization model while testing the first quantization model. If the test result of the first quantization model indicates that the first quantization model does not meet the quantization training requirements, it is equivalent to having started at least one round of iterative quantization training on the first quantization model in advance, and the usability is higher.

[0083] In S105, the model training device tests the second quantization model by simulating the hardware deployment environment to obtain a test result of the second quantization model.

[0084] The model training device may test the second quantization model by simulating the hardware deployment environment, to obtain the test result of the second quantization model. According to the test result of the second quantization model, the training strategy analysis of quantization training may be performed. For example, at least one round of iterative quantization training may be performed on the second quantization model again.

[0085] In some optional embodiments, such as illustrated in FIG. 5, after the test

result of the first quantization model is obtained in S102, the above method may further include the operation in S106.

[0086] In S106, a training strategy analysis result for quantization training of the neural network model is obtained based, at least in part, on the test result of the first quantization model.

[0087] Specifically, based on the test result of at least one test, it may be determined whether the current quantization training scheme of the neural network model is feasible, such as whether design of the loss function is reasonable, whether design of the quantization program is reasonable, etc., or the adjustment scheme of the current quantization training scheme may be determined, for example, whether it is necessary to modify one or more of the network structure, network hyperparameters, loss function or quantization scheme, etc., or adjustment strategies or other detailed information may be provided. In some optional examples, it may also be determined whether the current iterative quantization training needs to be stopped, for example, it is determined that early stop is needed, or it is determined that the neural network model has reached expectations, etc., which is not limited in the embodiments of the present disclosure.

[0088] In the embodiments of the present disclosure, the training strategy analysis result includes at least one of the following: terminating the quantization training on the neural network model, adjusting the quantization manner of at least one network layer in the neural network model, or adjusting the manner of the subsequent iterative quantization training of the neural network model.

[0089] In the embodiments of the present disclosure, training and testing are performed on the same platform, that is, on the same model training device, without model conversion and deployment, thereby optimizing the development process of the neural network model, and further, testing is performed during the training process, and the test result of the first quantization model may be obtained, such that the evaluation indexes may be more conveniently combined to determine whether it is necessary to continue the iterative quantization training of the neural network model. The evaluation indexes include, but are not limited to, an evaluation index used for such as early stop.

[0090] For example, when the test result of the first quantization model indicates that the first quantization model does not meet the quantization training requirements, the quantization training of the neural network model cannot be terminated. For another example, when the test result of the first quantization model indicates that the first quantization model is over-fitted, the quantization training of the neural network model may be terminated in advance according to the evaluation index.

[0091] According to the test result of the first quantization model, the model training device may also adjust the quantization manner for at least one network layer in the neural network model, for example, Post-quantization was previously used, and the quantization manner may be adjusted to Quantization-aware training according to the test result of the first quantization model.

[0092] According to the test result of the first quantization model, the model training device may also adjust the manner of the subsequent iterative quantization training of the neural network model. The quantization training manner includes, but is not limited to, the number of iterative quantization training, adjustment of the loss function, etc. For example, N rounds of iterative quantization training is performed on the neural network model to obtain the first quantization model, and M rounds of iterative quantization training may be performed on the first quantization model, where M and N are both positive integers and may be equal or unequal.

[0093] In the above embodiments, according to the test result of the first quantization model, the training strategy analysis result of the quantization training on the neural network model may be obtained. Based on the analysis result, the quantization training may be adjusted to make the quantization training of the neural network model more reasonable.

[0094] In some optional embodiments, the model training device may stop the iterative quantization training based on different training opportunities, so as to obtain the first quantization model.

[0095] In an optional implementation, the model training device may perform at least one round of iterative quantization training on the neural network model for a preset

number of times, test the obtained first quantization model, and obtain the test result of the first quantization model. The preset number of times may be much less than the previously set total number of iterative quantization training of the neural network model. For example, the total number of times is 1000, the preset number of times may be any positive integer less than 1000. According to the test result of the first quantization model in combined with an evaluation index such as early stop, the iterative quantization training of the neural network model may be terminated in advance, so as to avoid overfitting in the final quantization model.

[0096] In another possible implementation, the model training device may obtain the first quantization model when it is determined based on the preset test strategy that the first quantization model satisfies the test conditions, so as to test the first quantization model. The test conditions include, but are not limited to, change in the loss function corresponding to the neural network model is small, or precision of the neural network model is much less than a preset precision requirement, or the preset total number of iterative quantization has been reached.

[0097] In the above embodiments, the model training device may test the obtained first quantization model based on the training opportunity, thereby obtaining the test result of the first quantization model. In this way, it is beneficial to improve the development efficiency of the neural network model.

[0098] In some optional embodiments, the present disclosure also provides an image processing method, which may input an image to be processed into a quantization model to obtain an image processing result output by the quantization model.

[0099] The quantization model is a quantization model obtained by performing at least one round of iterative quantization training on the neural network model according to the quantization training method described above.

[00100] The image to be processed may be an image collected in a visual task. A visual task analysis may be performed on the collected image through the quantization model. The visual task analysis includes, but is not limited to, image classification, image semantic segmentation, human body key point detection or the like, and the availability is

high.

[00101] Corresponding to the above method embodiments, the present disclosure further provides device embodiments.

[00102] As illustrated in FIG. 6, FIG. 6 illustrates a block diagram of a quantization training apparatus according to an exemplary embodiment of the present disclosure. The apparatus includes a first quantization training module 210 and a first test module 220. The first quantization training module 210 is configured to perform at least one round of iterative quantization training on a neural network model by a model training device to obtain a first quantization model. The first test module 220 is configured to test the first quantization model through simulating a hardware deployment environment by the model training device to obtain a test result of the first quantization model.

[00103] In some optional embodiments, the first test module may include a first test submodule. The first test submodule is configured to test the first quantization model through invoking a target function by the model training device to obtain the test result of the first quantization model. The target function is used for simulating the hardware deployment environment.

[00104] In some optional embodiments, the apparatus may further include a model conversion module. The model conversion module is configured to perform conversion processing on the first quantization model to obtain a first test model. The conversion processing includes removing at least one target unit of the first quantization model, and the target unit is configured to perform quantization and/or dequantization on at least one of output data or network parameters of a network layer of the neural network model. The first test module may include a second test submodule, and the second test submodule is configured to test the first test model through simulating the hardware deployment environment by the model training device to obtain the test result of the first quantization model.

[00105] In some optional embodiments, the conversion processing is implemented by a target function used for simulating the hardware deployment environment.

[00106] In some optional embodiments, the first test module includes a third test

submodule. The third test submodule is configured to test the first quantization model by the model training device through simulating the hardware deployment environment and using fixed-point data obtained by fixed-point processing on test samples and network parameters of the first quantization model, to obtain the test result of the first quantization model.

[00107] In some optional embodiments, the apparatus further includes a second quantization training module and a second test module. The second quantization training module is configured to perform at least one round of iterative quantization training on the first quantization model by the model training device to obtain a second quantization model. The second test module is configured to test the second quantization model through simulating the hardware deployment environment by the model training device to obtain a test result of the second quantization model.

[00108] In some optional embodiments, the apparatus further includes a determination module. The determination module is configured to obtain a training strategy analysis result for quantization training of the neural network model based at least in part on the test result of the first quantization model. The training strategy analysis result includes at least one of the following: terminating the quantization training on the neural network model, adjusting a quantization manner of at least one network layer in the neural network model, or adjusting a manner of subsequent iterative quantization training of the neural network model.

[00109] In some optional embodiments, the second quantization training module includes a quantization training submodule. The quantization training submodule is configured to perform at least one round of iterative quantization training on the first quantization model in parallel with the process of performing test on the first quantization model by the model training device to obtain the second quantization model.

[00110] In some optional embodiments, the first test module includes any one of a fourth test submodule and a fifth test submodule. The fourth test submodule is configured to test the obtained first quantization model responsive to that at least one round of iterative quantization training on the neural network model reaches a preset number of times, to obtain the test result of the first quantization model. The fifth test submodule is

configured to test the first quantization model responsive to determining that the first quantization model satisfies a test condition based on a preset test strategy, to obtain the test result of the first quantization model.

[00111] The present disclosure further provides an image processing apparatus. The apparatus includes an image processing module, configured to input an image to be processed into a quantization model to obtain an image processing result output by the quantization model. The quantization model is a quantization model obtained according to the quantization training method described above.

[00112] The apparatus embodiment basically corresponds to the method embodiment, and the related details thereof may make reference to the part of the description of the method embodiment. The apparatus embodiment described above is merely illustrative. The units described as separate parts may or may not be physically separated, and parts displayed as units may or may not be physical units, and namely may be located in one place, or may also be distributed to a plurality of network units. Part or all of the modules may be selected to achieve the objectives of the solutions of the present disclosure according to a practical requirement. The disclosure may be understood and implemented by those of ordinary skill in the art without creative effort.

[00113] The embodiments of the present disclosure further provide a computer-readable storage medium having stored a computer program. The computer program is used to execute the quantization training method or the image processing method described above.

[00114] In some optional embodiments, the embodiments of the present disclosure provide a computer program product, including a computer-readable code. When the computer-readable code runs on a device, a processor in the device executes one or more instructions for implementing the quantization training method or the image processing method according to any of the above embodiments.

[00115] The computer program product may be implemented by hardware, software, or a combination thereof. In an optional embodiment, the computer program product is specifically embodied as a computer storage medium. In another optional embodiment,

the computer program product is specifically embodied as a software product, such as a software development kit (SDK), etc.

[00116] The embodiments of the present disclosure further provide a quantization training apparatus. The apparatus includes a processor and a memory for storing an executable instruction for the processor. The processor is configured to invoke the executable instruction stored in the memory to implement the quantization training method described above.

[00117] FIG. 7 illustrates a schematic diagram of a hardware structure of a quantization training apparatus according to an embodiment of the present disclosure. The quantization training apparatus 310 of the neural network model includes a processor 311, and may also include an input device 312, an output device 313, and a memory 314. The input device 312, the output device 313, the memory 314 and the processor 311 are connected to each other through a bus.

[00118] The memory includes, but is not limited to, a random access memory (RAM), a read-only memory (ROM), an erasable programmable read only memory (EPROM), or a compact disc read-only memory (CD-ROM). The memory is used for storing related instructions and data.

[00119] The input device is configured to input data and/or signals, and the output device is configured to output data and/or signals. The output device and the input device may be separated devices, and may also be a whole device.

[00120] The processor may include one or more processors, for example, including one or more central processing units (CPUs). In the case that the processor is one CPU, the CPU may be a single-core CPU, and may also be a multi-core CPU.

[00121] The memory is configured to store a program code and data of the network device.

[00122] The processor is configured to invoke the program code and data in the memory to execute the operations in the above method embodiments. The details may make reference to the description in the method embodiments, which will not be

elaborated herein.

[00123] It is to be understood that FIG. 7 only illustrates a simplified design of a quantization training apparatus. In a practical application, the quantization training apparatus may also include other necessary components, including but not limited to any number of input/output devices, processors, controllers, memories, etc., and all quantization training apparatuses that may implement the embodiments of the present disclosure are within the protection scope of the present disclosure.

[00124] The embodiments of the present disclosure further provide an image processing apparatus. The apparatus includes a processor and a memory for storing an executable instruction for the processor. The processor is configured to invoke the executable instruction stored in the memory to implement the image processing method described above.

[00125] Other implementation schemes of the present disclosure may be conceived easily by those skilled in the art upon consideration of the specification and practice of the invention disclosed herein. The present disclosure is intended to cover any variations, uses, or adaptive changes of the present disclosure. These variations, uses, or adaptive changes follow the general principles of the present disclosure and include common knowledge or commonly-used technical means in the art that are not disclosed in the present disclosure. The specification and the embodiments are exemplary only, and the true scope and spirit of the present disclosure are subject to the following claims.

[00126] The above are only the preferred embodiments of the present disclosure and not intended to limit the present disclosure. Any modification, equivalent replacement, improvement, etc. within the spirit and principle of the present disclosure shall fall within the protection scope of the present disclosure.

CLAIMS

1. A quantization training method, comprising:

performing at least one round of iterative quantization training on a neural network model to obtain a first quantization model; and

testing the first quantization model by simulating a hardware deployment environment to obtain a test result of the first quantization model.

2. The method according to claim 1, wherein testing the first quantization model by simulating the hardware deployment environment to obtain the test result of the first quantization model comprises:

testing the first quantization model by invoking a target function to obtain the test result of the first quantization model, wherein the target function is used to simulate the hardware deployment environment.

3. The method according to claim 1 or 2, further comprising:

performing conversion processing on the first quantization model to obtain a first test model, wherein the conversion processing comprises removing at least one target unit of the first quantization model, and the target unit is configured to perform quantization and/or dequantization on at least one of output data or network parameters of a network layer in the neural network model;

testing the first quantization model by simulating the hardware deployment environment to obtain the test result of the first quantization model comprises:

testing the first test model by simulating the hardware deployment environment to obtain the test result of the first quantization model.

4. The method according to claim 3, wherein the conversion processing is implemented by a target function used for simulating the hardware deployment environment.

5. The method according to any one of claims 1 to 4, wherein testing the first quantization model by simulating the hardware deployment environment to obtain the test result of the first quantization model comprises:

testing the first quantization model by simulating the hardware deployment environment and using fixed-point data obtained by fixed-point processing on test samples and network parameters of the first quantization model to obtain the test result of the first quantization model.

6. The method according to any one of claims 1-5, further comprising:

performing at least one round of iterative quantization training on the first quantization model to obtain a second quantization model; and

testing the second quantization model by simulating the hardware deployment environment to obtain a test result of the second quantization model.

7. The method according to any one of claims 1-6, further comprising:

obtaining a training strategy analysis result for quantization training of the neural network model based on the test result of the first quantization model, wherein the training strategy analysis result comprises at least one of the following: terminating the quantization training on the neural network model, adjusting a quantization manner of at least one network layer in the neural network model, or adjusting a manner of the subsequent iterative quantization training of the neural network model.

8. The method according to claim 6 or 7, wherein performing at least one round of iterative quantization training on the first quantization model to obtain the second quantization model comprises:

performing at least one round of iterative quantization training on the first quantization model in parallel with the process of testing the first quantization model to obtain the second quantization model.

9. The method according to any one of claims 1-8, wherein testing the first quantization model to obtain the test result of the first quantization model comprises any one of the following:

testing the obtained first quantization model responsive to that at least one round of iterative quantization training on the neural network model reaches a preset number of times to obtain the test result of the first quantization model; or

testing the first quantization model responsive to determining that the first quantization model satisfies a test condition based on a preset test strategy to obtain the test result of the first quantization model.

10. An image processing method, comprising:

inputting an image to be processed into a first quantization model to obtain an image processing result output by the first quantization model, wherein the first quantization model is a quantization model obtained by the method according to any one of claims 1-9.

11. A quantization training apparatus, comprising:

a first quantization training module, configured to perform at least one round of iterative quantization training on a neural network model to obtain a first quantization model; and

a first test module, configured to test the first quantization model by simulating a hardware deployment environment to obtain a test result of the first quantization model.

12. An image processing apparatus, comprising:

an image processing module, configured to input an image to be processed into a quantization model to obtain an image processing result output by the quantization model, wherein the quantization model is a quantization model trained by the method according to any one of claims 1-9.

13. A computer-readable storage medium, having stored a computer program that is configured to execute the quantization training method according to any one of claims 1-9 or the image processing method according to claim 10.

14. A quantization training apparatus, comprising:

a processor; and

a memory for storing an executable instruction for the processor;

wherein the processor is configured to invoke the executable instruction stored in the memory to implement the quantization training method according to any one of claims 1-9.

15. An image processing apparatus, comprising:

a processor; and

a memory for storing an executable instruction for the processor;

wherein the processor is configured to invoke the executable instruction stored in the memory to implement the image processing method according to claim 10.