Pre-training Service System, and Service Providing Method Based on Pre-training Service System

The present disclosure claims the priority to Chinese Patent Application No. 202210353465.X, filed with the China National Intellectual Property Administration on April 06, 2022 and entitled "Pre-training service system, and service providing method based on pre-training service system". Contents of the present disclosure are hereby incorporated by reference in entirety of the Chinese Patent Application.

Technical Field

At least one embodiment of the present disclosure relates to the field of deep learning, and in particular to a pre-training service system, and a service providing method based on a pre-training service system.

Background

With the rapid development of artificial intelligence (AI) technology, a variety of large-scale pre-trained models have emerged, and capacities (e.g., comprehension, generation, and retrieval) of the pre-trained models have been applied to various industries with remarkable success. However, since the barrier to entry for using the pre-trained models is high, an urgent issue of reducing the difficulty in use and providing a platform for using the pre-trained models through simple operations for more users has not been solved yet. Moreover, since the pre-trained models have strong understanding, generation and retrieval capacities, avoiding the abuse of the pre-trained models should also be taken into consideration.

Summary

In view of the above, at least one embodiment of the present disclosure provides a pre-training service system.

To realize the above objective, the technical solutions provided in at least one embodiment of the present disclosure are as follows.

In a first aspect, a pre-training service system is provided in at least one embodiment of the present disclosure. A model maintained by the pre-training service system includes obfuscated model structure codes and obfuscated model parameters, and the obfuscated model structure codes and the obfuscated model parameters are generated after an obfuscation processing is pre-performed on original model structure codes and original model parameters respectively; and

the pre-training service system includes:

a producer service module arranged for providing a model pre-training process for a pre-training dataset for a model producer to generate a pre-trained model;

an optimizer service module arranged for optimizing the pre-trained model according to a fine-tuning dataset provided by a model optimizer to obtain an optimized model; and

a consumer service module arranged for providing a service interface for at least one of the pre-trained model and/or the optimized model for a model consumer, the pre-trained model or the optimized model being arranged for performing inference on data provided by the model consumer to output a model prediction result.

In a second aspect, a service providing method based on a pre-training service system is provided in at least one embodiment of the present disclosure. The pre-training service system includes a producer service module, an optimizer service module, and a consumer service module; a model maintained by the pre-training service system includes obfuscated model structure codes and obfuscated model parameters, and the obfuscated model structure codes and the obfuscated model parameters are generated after an obfuscation processing is pre-performed on original model structure codes and original model parameters respectively; and the method includes:

acquiring identity information of a target user accessing the pre-training service system, so as to determine a preset user type to which the target user belongs; and

opening a service module matching the preset user type, to which the target user belongs, to the target user; where the producer service module is opened to the target user in response to the target user being a model producer, where the producer service module is arranged for providing a model pre-training process for a pre-training dataset for the model producer to generate a pre-trained model;

the optimizer service module is opened to the target user in response to the target user being a model optimizer, where the optimizer service module is arranged for optimizing the pre-trained model according to a fine-tuning dataset provided by the model optimizer to obtain an optimized model; and

the consumer service module is opened to the target user in response to the target user being a model consumer, where the consumer service module is arranged for providing a service interface for at least one of the pre-trained model and the optimized model to the model consumer, and the pre-trained model or the optimized model being arranged for performing inference on data

provided by the model consumer to output a model prediction result.

In a third aspect, a service providing apparatus based on a pre-training service system is provided in at least one embodiment of the present disclosure. The pre-training service system includes a producer service module, an optimizer service module, and a consumer service module; a model maintained by the pre-training service system includes obfuscated model structure codes and obfuscated model parameters, and the obfuscated model structure codes and the obfuscated model parameters are generated after an obfuscation processing is pre-performed on original model structure codes and original model parameters respectively; and the apparatus includes:

an acquisition unit arranged for acquiring identity information of a target user accessing the pre-training service system, so as to determine a preset user type to which the target user belongs; and

an opening unit arranged for opening a service module matching the preset user type to which the target user belongs to the target user; where

the producer service module is opened to the target user in response to the target user being a model producer, where the producer service module is arranged for providing a model pre-training process for a pre-training dataset for the model producer to generate a pre-trained model;

the optimizer service module is opened to the target user in response to the target user being a model optimizer, where the optimizer service module is arranged for optimizing the pre-trained model according to a fine-tuning dataset provided by the model optimizer to obtain an optimized model; and

the consumer service module is opened to the target user in response to the target user being a model consumer, where the consumer service module is arranged for providing a service interface for at least one of the pre-trained model and/or the optimized model to the model consumer, and the pre-trained model or the optimized model being arranged for performing inference on data provided by the model consumer to output a model prediction result.

In a fourth aspect, an electronic device is provided in at least one embodiment of the present disclosure. The electronic device includes a processor, a memory, and a non-transitory storage memory; where

the processor is arranged for reading a computer program from the non-transitory storage memory into the memory, and executing the method in the second aspect based on the computer program.

In a fifth aspect, a computer-readable storage medium is provided in at least one embodiment of the present disclosure. The computer-readable storage medium stores a computer program, where when executed by a processor, the computer program implements the method in the second aspect.

In a sixth aspect, a computer program product is provided in at least one embodiment of the present disclosure. The computer-readable storage medium includes a computer program, where when the computer program runs, the computer program causes a computer to execute the method in the second aspect.

In the technical solutions provided by the present disclosure, different types of services are provided for users having different identities through the pre-training service system including the producer service module, the optimizer service module, and the consumer service module. Accordingly, the difficulty of using the pre-training model is reduced, the universality of the pre-trained model is improved, and the model optimizer having no capacity to generate the pre-trained model is endowed with a capacity to rapidly and conveniently obtain the optimized model based on the pre-trained model. Moreover, the service interface is also provided to the model consumer, so that the model consumer can rapidly perform inference based on the obtained pre-trained model or optimized model through the service interface to obtain the model prediction result. Moreover, since the model structure codes and the model parameters provided by the pre-training service system are the obfuscated model structure codes and the obfuscated model parameters respectively, which are functionally equivalent to the original model structure codes and the original model parameters, but it is difficult to reproduce the pre-trained model through the obfuscated model structure codes, the abuse of the pre-trained model can be avoided based on the obfuscated structure codes and the obfuscated model parameters, and the effect of protecting the pre-training model can be reached.

Brief Description of the Drawings

FIG. 1 is a schematic structural diagram of a pre-training service system according to some embodiments of the present disclosure.

FIG. 2 is a schematic flowchart of a service providing method based on a pre-training service system according to some embodiments of the present disclosure.

FIG. 3 is a schematic structural diagram of an electronic device according to some embodiments of the present disclosure.

FIG. 4 is a schematic diagram of a service providing apparatus based on a pre-training service

system according to some embodiments of the present disclosure.

Detailed Description

Illustrative embodiments will be described in detail herein and illustratively shown in the accompanying drawings. When the following description relates to the accompanying drawings, the same numbers in different accompanying drawings denote the same or similar elements, unless indicated otherwise. Implementations described in the following illustrative embodiments do not denote all implementations consistent with at least one embodiment of the present disclosure. On the contrary, the implementations are instances of an apparatus and method consistent with some aspects of at least one embodiment of the present disclosure as recited in the appended claims.

It should be noted that in other embodiments, steps of corresponding methods may not necessarily be executed in the order shown and described in the present disclosure. In some other embodiments, the method may include more or fewer steps than those described in the present disclosure. In addition, a single step described in the present disclosure may be decomposed into multiple steps for description in other examples, and multiple steps described in the present disclosure may also be combined into a single step for description in other embodiments.

The complication of application scenarios of the AI technologies leads to differences in task objectives, training data, problem settings, etc. in different application scenarios. Therefore, through providing a user with a general pre-trained model, the user may finely tune this pre-trained model, so as to obtain an application mode of an optimized model that satisfies a specific application scenario. Accordingly, the demand for training data is reduced, the training efficiency is improved, an increased quantity of small and medium-sized enterprises have the capacity to obtain the optimized model that satisfies the specific application scenario, and the application and popularization of the AI technology are promoted. Moreover, the user may also select different pre-trained models or optimized models serving specific scenarios according to the application scenarios, and perform inference based on data of the user, so as to obtain model output results based on the pre-trained models or optimized models selected.

To facilitate providing different types of services based on the pre-trained model for the users having different identity types, a pre-training service system shown in FIG. 1 is provided in some embodiments of the present disclosure. The pre-training service system includes a producer service module 11, an optimizer service module 12, and a consumer service module 13. The producer service module 11 is arranged for providing a model pre-training process for a pre-training dataset 111 for a model producer to generate a pre-trained model 112. The optimizer service module 12 is arranged for optimizing the pre-trained model 112 according to a fine-tuning dataset 121 provided

by a model optimizer to obtain an optimized model 122. The consumer service module 13 is arranged for providing a service interface for at least one of the pre-trained model 112 and the optimized model 122 for a model consumer, where the pre-trained model 112 or the optimized model 122 is arranged for performing inference on data 131 provided by the model consumer to output a model prediction result 132.

The above model producer refers to an algorithm developer on a pre-training service system side, and the model producer is a producer of the pre-trained model 112. In some embodiments of the present disclosure, the model producer activates distributed pre-training through the model pre-training function provided by the pre-training service system based on the pre-training dataset 111, and produces the pre-trained model 112. The pre-trained model 112 is the basis for the pre-training service system to provide an artificial intelligence-related service. Other pre-trained models are obtained after the pre-trained model 112 is modified.

In some embodiments of the present disclosure, the model pre-training function implements an activation checkpointing technology in a pre-training process of the pre-trained model 112. Several checkpoints are configured in a network, and some temporarily-useless intermediate results are discarded in a forward process to reduce memory occupation. If these intermediate results are needed in a subsequent process, these intermediate results are recovered from the nearest checkpoint through additional forward calculation. In this way, a video memory is saved on, and a tedious process of calculating from scratch is avoided, so that occupation of the video memory caused by intermediate activation during training of the pre-trained model is reduced.

In addition, a mixed precision training technology is also employed in the pre-training process. The training of the pre-trained model is deeply learnt through the 16-bit floating point (FP16), so that the video memory required for training is reduced, and a calculation speed is increased. Moreover, the pre-training system also improves the calculation efficiency in the pre-training process based on sparse expert parallelism and data parallelism technologies, ensuring low carbon and environmental protection of the pre-training process. The sparse expert parallelism technology indicates that a sparsely-gated mixture of experts (MoE) network splits a large model into multiple small models. Independent parameters are retained in each small model. In the model pre-training process, each sample is calculated through an appropriate small model and parameters corresponding to this model, which are selected according to a preset strategy, instead of being calculated through all the small models. Although a total number of parameters for pre-training in such a solution is extremely large, a small portion of the video memory is activated, which reduces memory usage during calculation.

In some embodiments of the present disclosure, to make the pre-trained model 112 have a more diversified capacity and higher versatility, a multi-modality pre-training form is employed in the pre-training process of the pre-trained model 112. The modality refers to an existing form of data, such as a file format including text, audio, images, and videos.

In some embodiments of the present disclosure, in the pre-training process of the pre-trained model 112, the pre-trained model 112 may use multiple tasks including language modeling, text denoising, image captioning, image-based text denoising, and etc. In addition, in the pre-training process, a size and field breadth of the pre-training dataset 111 affect the capacity and versatility of the pre-trained model 112. Therefore, in some embodiments of the present disclosure, ultra-large-scale single-modality and multi-modality datasets covering multiple fields are used as the pre-training dataset 111 of the pre-trained model 112 in the pre-training process of the pre-trained model 112. Through activating multi-modality pre-training and expanding the size and coverage field of the pre-training dataset 111, the pre-trained model 112 may have the single-modality understanding and generation capacity and the multi-modality understanding and generation capacity, and have favorable adaptability to tasks in different fields. Therefore, the versatility of the pre-trained model 112 can be improved.

Corresponding to the model developer, the model optimizer refers to a user in another field who intendeds to obtain the optimized model 122 adapted to the field to which the user belongs based on the pre-trained model 112. In some embodiments of the present disclosure, the pre-training service system shown in FIG. 1 provides the optimizer service module 12 for the service optimizer. The above optimizer service module 12 optimizes the pre-trained model 112 produced by the producer service module 11 according to the fine-tuning dataset 121 provided by the model optimizer to obtain the optimized model 122.

In some embodiments of the present disclosure, the model pre-training function provided by the pre-training service system is arranged for performing training a general pre-trained model according to the pre-training dataset 111. Then, the producer service module generates a dedicated pre-trained model according to the general pre-trained model and the dedicated pre-trained model corresponds to a downstream task instructed by the model producer. In other words, the model producer indicates the pre-training service system according to an atomic capacity of the general pre-trained model to generate the dedicated pre-trained model having a special purpose and suitable for a type of downstream task based on the general pre-trained model. The atomic capacity of the general pre-trained model includes a cross-modality generation capacity, a cross-modality comprehension capability capacity, a cross-modality retrieval capacity, etc. Correspondingly, the dedicated pre-trained model includes a cross-modality generation model, a cross-modality

comprehension capability model, a cross-modality retrieval model, etc. The pre-trained model 112 includes the general pre-trained model and the dedicated pre-trained model.

In some embodiments of the present disclosure, the producer service module 11 contained in the pre-training service system is further arranged for generating a code development template corresponding to the dedicated pre-trained model. The pre-training service system provides the code development template corresponding to the dedicated pre-trained model to the model optimizer through the optimizer service module 12.

Any pre-trained model primarily includes model structure codes and model parameters. The model structure codes refer to codes for constructing the pre-trained model. The model may be deemed as an assumption, and the model parameters refer to a specific adjustment to the hypothesis according to a specific dataset, and values of the model parameters may define a model function.

With reference to the above code development template, the model optimizer issues a model modification instruction to the pre-training service system according to optimization demands of the model optimizer. Specifically, the optimizer service module 12 contained in the pre-training service system provides a model modification interface to the model optimizer. With reference to the code development template, the above model optimizer issues the model modification instruction for the model structure codes and/or the model parameters of the dedicated pre-trained model through the model modification interface. The optimizer service module modifies the model structure codes and/or model parameters according to the model modification instruction. Then, training is activated for a modified pre-trained model according to the fine-tuning dataset 121 provided by the model optimizer to obtain the optimized model.

In some embodiments of the present disclosure, the model optimizer who intends to rapidly try out the pre-trained model may activate, based on data (fine-tuning dataset) in the field, a multi-machine multi-card distributed training task through the pre-training service system to produce optimized model parameters belonging to the model optimizer, through modifying illustrative parameters provided by the above code development template. The optimized model is produced by combining the optimized model parameters with the model structure codes of the pre-trained model.

In some illustrative embodiments of the present disclosure, if the model optimizer is intended to further optimize the pre-trained model, the pre-training service system may further provide an instance of optimizing the pre-trained model and applying the optimized model to other fields. With reference to the above instance, the optimizer may issue a modification instruction to the

pre-trained model according to demands of the optimizer, and input the fine-tuning dataset of the field. The pre-training service system activates the distributed training task to generate optimized model parameters, so as to obtain the optimized model.

The optimized model is trained based on the fine-tuning dataset provided by the model optimizer. The model optimizer may optimize and finely tune the model according to demands of the model optimizer. Therefore, the optimized model is suitable for the field to which the model optimizer belongs. In addition, in a process for performing model optimization by the pre-training service system, the model optimizer may modify the model structure codes or model parameters according to the code development template provided by the pre-training service system, and activate model training to obtain the optimized model suitable for the industry to which the model optimizer belongs, instead of focusing on specific implementation details of a development process of the pre-trained model, distributed training of the pre-trained model, an evaluation flow of the pre-trained model, etc., for example, data input/output (IO) and data parallelism. The pre-training service system reduces the difficulty of optimizing the pre-trained model by the user, so that the model optimizer who has no capacity to develop the pre-trained model may also obtain the optimized model through the pre-training service system.

In some embodiments of the present disclosure, the pre-training service system may provide industry solutions (such as copy generation, search recall, clothing design, intelligent conversation, financial questions and answers, etc.) that are towards maturity for other users (including other developers) as instances for reference. The model optimizer may also modify the model with reference to the industry solutions, except for transmitting the modification instruction to the pre-trained model with reference to the code development template.

In some embodiments of the present disclosure, the pre-training service system further provides a sharing interface. The model optimizer may upload the optimization process of the pre-trained model to the pre-training service system through the sharing interface for sharing the optimization process as a new instance, and provide the optimization process for other model optimizers as an optimization reference. An application scope of the pre-trained model can be expanded through a case sharing function, and the difficulty of using the pre-training service system can be further reduced.

In the present disclosure, the user of the pre-training service system further includes the model consumer. The model consumer is a user who intendeds to perform inference on data of the model consumer based on the pre-trained model or the optimized model to obtain a model prediction result. The pre-training service system provides a service interface for at least one of the pre-trained

model and the optimized model to the model consumer through the consumer service module, where the pre-trained model or the optimized model is arranged for performing inference on data provided by the model consumer to output the model prediction result. As shown in FIG. 1, the model consumer may invoke the pre-trained model 112 or the optimized model 122 for inference. The above inference process may be offline. The pre-training service system deploys the pre-trained model 112 or the optimized model 122 selected by the model consumer in a device corresponding to the model consumer through the service interface, for performing offline inference on the data provided by the model consumer. Certainly, the above pre-trained model 112 or the above optimized model 122 may also be deployed as a low-latency online service. The model consumer directly inputs data 131 of the model consumer into a web page provided by the pre-training service system. The web page is equivalent to the service interface provided by the pre-training service system. After the data 131 of the model consumer are obtained, the pre-trained model 112 or optimized model 122 is invoked for online inference, and the model prediction result is obtained and provided to the model consumer via the web page.

In some embodiments of the present disclosure, the pre-training service system may provide various different single-modality or cross-modality artificial intelligence services for the user based on the cross-modality generation, cross-modality retrieval, and cross-modality comprehension capability capacities of the pre-trained model.

For example, a news event may be displayed in various forms such as text, images, and videos. Therefore, information content obtained through performing information retrieval may be seriously limited when retrieving single-modality data, and data in other modalities are required to enrich the cognition of the same thing or event. In this case, cross-modality retrieval is required to implement retrieval between data in different modalities. In some embodiments of the present disclosure, the pre-training service system may obtain a video of the news event according to news information in a form of a picture input by the user based on the cross-modality retrieval capacity of the pre-trained model.

In some embodiments of the present disclosure, the pre-training service system may provide a clothing design service to the model consumer based on the cross-modality generation capacity of the pre-trained model. For example, the model consumer may input “men’s striped shirt” via the web page provided by the pre-training service system to generate several clothing design pictures conforming to the text description “men’s striped shirt” through calculation of a target pre-trained model in the field of clothing design provided by the pre-training service system. The cross-modality generation capacity of the pre-trained model is used in the above instance. Certainly, although providing the cross-modality artificial intelligence service for the user, the pre-training service

system also has the capacity to provide the single-modality artificial intelligence service. For example, when performing a clothing design operation as follows through an offline inference function of the target pre-trained model in the field of clothing design provided by the pre-training service system, the model consumer may input two clothing pictures in different styles to obtain an offline inference result, namely a clothing picture obtained by merging two pieces of clothing in different styles after offline inference.

Except for the clothing design function, the pre-training service system may also perform text processing-related training inference, such as copy generation, based on the cross-modality retrieval, generation, and understanding capacities of the pre-trained model. For example, in the inference process of the target pre-training model in the field of text processing, a picture of a latex pillow is input. When inference is performed based on the target pre-trained model in the field of text processing provided by the pre-training service system, a propaganda copy for the latex pillow may be obtained based on the cross-modality comprehension capability capacity of the pre-trained model. In addition, the pre-training service system also implements conversion from text to text based on the above capacities of the pre-trained model, and performs text denoising and language modeling in the conversion process.

As can be seen from the above introduction of the pre-training service system, the pre-trained model plays a significant role in the service process of the pre-training service system. The user may invoke the target pre-trained model through the pre-training service system. Since the pre-trained model has strong text and image generation capacity, to avoid abuse of the pre-trained model, the pre-training service system is further provided with various defense mechanisms to prevent the pre-trained model from being abused. The defense mechanism is also suitable for avoiding abuse of the optimized model.

As can be seen from the above introduction, the pre-trained model or the optimized model includes the model structure codes and the model parameters. For the model structure codes, according to a common framework in the industry, etc., the user may reversely reproduce the complete model structure codes to obtain the pre-trained model or the optimized model by combining the reproduced model structure codes with the model parameters obtained by other means. In this case, it is not conducive to control of the application of the pre-trained model or the optimized model. Since the model parameters defines the model functions, the protection on the model parameters and the protection on the model structure codes are equally important for the pre-training service system. When the pre-trained model is protected as a whole, the protection on the model structure codes and the protection on the model parameters are required to be taken into consideration.

In some embodiments of the present disclosure, at least one of the pre-trained model and the optimized model includes obfuscated model structure codes and obfuscated model parameters, where the obfuscated model structure codes and the obfuscated model parameters are generated after an obfuscation processing is pre-performed on original model structure codes and original model parameters respectively. Through performing the obfuscation on the model structure codes and the model parameters, the obfuscated model structure codes and the obfuscated model parameters have equivalent functions to the original model structure codes and the original model parameters, but cannot be reproduced in form, so that the effect of avoiding abuse of the pre-trained model is reached. Values of the model parameters may define the model function.

Specifically, the obfuscation processing may be implemented in a form as follows. On the basis of performing a conventional obfuscation operation (generally by removing comments, adding irrelevant codes, renaming functions, classes, and variables, etc., so as to reduce code readability and increase a cracking threshold, while a running result is consistent after obfuscation with a running result before obfuscation) on at least a part of codes in the original model structure codes of the pre-trained model, at least one calculation process in the model structure code is modified. Moreover, some parameters, corresponding to the calculation process, of the model parameters are modified cooperatively. Therefore, a running result of an obfuscated pre-trained model formed by combining the obfuscated model structure codes and the obfuscated model parameters correspondingly is consistent with a running result of an original pre-trained model before obfuscation. Since the obfuscation process cooperates with modification and adjustment of the model parameters, it is difficult for others to locate a specific position where obfuscation occurs, and the above method further increase the difficulty of reproducing the model structure codes, and realizes the protection on the model structure codes.

The above step that at least one calculation process in the model structure code is modified includes the following steps. Forward obfuscation processing is performed on at least a part of codes in the original model structure codes, and reverse obfuscation processing is performed on model parameters, associated with some modified codes, in the original model parameters. Therefore, the difficulty of reproducing the pre-trained model is increased without changing the running result of the pre-trained model.

For example, the above forward obfuscation processing may include the following step. A certain vector is subtracted in the calculation process of the model structure codes on the basis of performing the conventional obfuscation operation on at least a part of codes in the original model structure codes of the pre-trained model. The reverse obfuscation processing may include the following step. The vector is added to a corresponding portion of the model parameters for

compensation. Therefore, the running result of the obfuscated pre-trained model is consistent with the running result of the original pre-trained model before obfuscation.

To further improve the security of the pre-trained model, in some embodiments of the present disclosure, the model parameters are stored in a storage space managed by the pre-training service system. When a target user invokes the target pre-trained model or the target optimized model, the corresponding model parameters are acquired from the storage space according to a reading path indicated in an invocation request from the target user. The storage space may be object storage (cloud storage) managed by the pre-training service system. Alternatively, the storage space may also be directly provided by the pre-training service system. The above model parameters may be the obfuscated model parameters. After obtaining the obfuscated model parameters, the pre-training service system combines the obfuscated model parameters with the obfuscated model structure codes to obtain the target pre-trained model or the target optimized model.

The users of the pre-training service system have different identities, including the model developer, the model optimizer, and the model consumer. Different types of users have different demands to complete different tasks based on the pre-training service system. Therefore, it is not conducive to the protection on the pre-trained model if all use permission is opened to all types of users on the pre-training service system.

To provide the model use permission according to the user identity, the pre-training service system in the present disclosure further provides the services as follows.

The optimizer service module is further arranged for identifying whether the model optimizer has a permission to invoke the obfuscated model parameters corresponding to the pre-trained model; and combining the obfuscated model parameters with the obfuscated model structure codes to obtain the pre-trained model and provide the pre-trained model to the model optimizer in response to the model optimizer having the permission to invoke the obfuscated model parameters; and/or,

the consumer service module is further arranged for identifying whether the model consumer has a permission to invoke the obfuscated model parameters corresponding to the pre-trained model or the optimized model; and combining the obfuscated model parameters with the obfuscated model structure codes to obtain the pre-trained model or the optimized model and provide the pre-trained model or the optimized model to the model consumer in response to the model consumer having the permission to invoke the obfuscated model parameters.

For example, the optimizer service module contained in the pre-training service system identifies whether the model optimizer has a permission to invoke the model parameters of the target pre-trained model selected by the model optimizer, where the model parameters may be the obfuscated model parameters. If the model optimizer has the permission to invoke the above obfuscated model parameters, after obtained by the pre-training service system, the obfuscated model parameter are combined with the corresponding obfuscated model structure codes to obtain the target pre-trained model. The target pre-trained model is provided for the above model optimizer, so as to optimize the target pre-trained model by the model optimizer. The model consumer has a similar identification process with the above model optimizer.

Certainly, in the running process of the pre-training service system, to invoke and use the pre-trained model or the optimized model, the pre-training service system also runs model access codes and model training inference codes for any pre-trained model or optimized model, except for the model structure codes and the model parameters of the pre-trained model and the optimized model, decouples these codes from the model structure codes, so as to obtain independent model access codes and model training inference codes, and encrypts the model access codes and the model structure codes. Accordingly, the exposure of the storage position of the pre-trained model or the optimized model or the leakage of the model structure codes can be avoided, the security of the pre-training service system can be improved, and abuse of the pre-trained model or the optimized model can be avoided. The model training inference codes may be maintained in plaintext by the pre-training service system.

A service providing method based on a pre-training service system shown in FIG. 2 is provided in some embodiments of the present disclosure. The method may be applied to the pre-training service system shown in FIG. 1. The above pre-training service system may include the producer service module 11, the optimizer service module 12, and the consumer service module 13. A model maintained by the above pre-training service system includes obfuscated model structure codes and obfuscated model parameters, where the obfuscated model structure codes and the obfuscated model parameters are generated after an obfuscation processing is pre-performed on original model structure codes and original model parameters respectively; and the method may include the following steps.

In step S201, identity information of a target user accessing the pre-training service system is acquired to determine a preset user type to which the target user belongs.

For example, the pre-training service system may determine the identity information of the target user according to a login account of the target user to determine the preset user type to which

the target user belongs. Alternatively, the pre-training service system may determine the identity information of the target user according to an invocation request from the target user.

In step S202, a service module matching the preset user type to which the target user belongs is opened to the target user.

The producer service module is opened to the target user in response to the target user being a model producer, where the producer service module is arranged for providing a model pre-training process for a pre-training dataset for the model producer to generate a pre-trained model.

The optimizer service module is opened to the target user in response to the target user being a model optimizer, where the optimizer service module is arranged for optimizing the pre-trained model according to a fine-tuning dataset provided by the model optimizer to obtain an optimized model.

The consumer service module is opened to the target user in response to the target user being a model consumer, where the consumer service module is arranged for providing a service interface for at least one of the pre-trained model and the optimized model to the model consumer, and the pre-trained model or the optimized model is arranged for performing inference on data provided by the model consumer to output a model prediction result.

The above model pre-training function is used for performing training, according to the pre-training dataset, a general pre-trained model; and the pre-trained model includes at least one of the general pre-trained model and a dedicated pre-trained model generated by the producer service module according to the general pre-trained model and the dedicated pre-trained model corresponds to a downstream task instructed by the model producer.

In some embodiments of the present disclosure, the above service providing method based on a pre-training service system further includes the following steps. A code development template corresponding to the dedicated pre-training model is generated. The code development template is provided to the model optimizer; and model structure codes and/or model parameters of the dedicated pre-trained model are modified according to a model modification instruction issued by the model optimizer.

For example, in some embodiments of the present disclosure, it is assumed that a model optimizer (a target user A) in the field of clothing design needs to perform modification based on a target pre-trained model, so as to obtain an optimized model for a target field, i.e. the field of clothing design, and that the target pre-trained model is the pre-trained model 112 shown in FIG. 1. In this case, after the target user A in the field of clothing design logs into the pre-training service

system, the pre-training system obtains identity information of the target user, and confirms that the target user is the model optimizer. The pre-training service system opens the optimizer service module 12 to the target user A. The target user A may modify parameters of the pre-training model 112 with reference to the code development template provided by the pre-training service system, and train a modified pre-trained model based on data in a fine-tuning dataset 121 of the target user A belonging to the field of clothing design, so as to obtain an optimized model 122. Alternatively, based on optimization permission granted by the pre-training service system, the target user A may also issue a modification instruction instructing the pre-training service system to modify the corresponding model structure codes through a modification interface provided by the pre-training service system, and similarly, activate model training based on the fine-tuning dataset, so as to obtain the optimized model 122. The optimized model 122 is suitable for the field of clothing design.

In some illustrative embodiments of the present disclosure, in response to determining that a user type of a target user B is the model consumer, the pre-training service system opens the consumer service module 13 to the target user B. The consumer service module 13 is arranged for providing a service interface for at least one of the pre-trained model and the optimized model to the model consumer, where the pre-trained model or the optimized model is arranged for performing inference on data provided by the model consumer to output a model prediction result. For example, when intended to obtain a clothing design solution of a particular theme through an artificial intelligence service provided by the pre-training service system, the target user B requests to invoke a target optimized model in the field of clothing design, i.e. the optimized model 122 obtained by optimizing on the basis of the pre-trained model 112 by the developer, the target user A, in the previous embodiment. The consumer service module 13 contained in the pre-training service system provides model use permission to perform inference based on the optimized model 122 to obtain the model prediction result for the target user B through the service interface. Specifically, the target user B may input “men’s striped shirt” based on an inference service deployed online in the pre-training service system, and invoke the optimized model 122 to perform online inference calculation to obtain a model prediction result, for example, several clothing design pictures conforming to the text description “men’s striped shirt”. Certainly, the inference function may also be set as an offline function. For example, the optimized model 122 is deployed in a device corresponding to the model consumer (namely the target user B) through the service interface provided by the consumer service module contained in the pre-training service system, for performing offline inference on the data provided by the model consumer.

Similarly, the model consumer may also invoke optimized models in different fields obtained by model optimizers in the field of text processing, finance, etc. through the same optimization method

as that in the above example, and perform inference, so as to obtain model prediction results in different fields.

Since the model optimizer may provide the optimized model for other users for use after optimizing the target pre-trained model through the pre-training service system. In some embodiments of the present disclosure, the pre-training service system further provides the following function of storing parameters of the optimized model produced by the pre-training service system into a storage space managed by the pre-training service system. The uploaded parameters of the optimized model may be invoked according to a reading path provided by the target user through the above method. For example, as in the above embodiment, after obtaining the optimized model 122, the target user A stores the model parameters into the storage space managed by the pre-training service system. When the target user B is intended to invoke the optimized model 122, the pre-training service system may invoke the model parameters of the optimized model 122 according to the reading path provided by the user B. Certainly, the above model parameters may be the obfuscated model parameters, and the optimized model 122 may be obtained by combining the obfuscated model parameters with the obfuscated model structure codes. The above obfuscated model structure codes and obfuscated model parameters are generated after an obfuscation processing is pre-performed on original model structure codes and original model parameters of the optimized model 122 respectively. Reference may be made to the above relevant introduction of the pre-training service system for a specific method for the above an obfuscation processing.

In some embodiments of the present disclosure, the pre-training service system may identify whether the target user has a permission to invoke the obfuscated model parameters, and combine the obfuscated model parameters with the obfuscated model structure codes to obtain the target pre-trained model or the target optimized model in response to the target user having the permission to invoke the obfuscated model parameters. The model parameters are generally stored in a form of a ckpt file. To ensure that the ckpt file of the model parameters cannot be used independently from the pre-training service system, a license server may be introduced to identify whether the target user has the permission to invoke the model parameters (the ckpt file) provided by the pre-training service system. The model parameters are combined with the model structure codes to obtain the target pre-trained model in response to the target user having the above permission.

FIG. 3 is a schematic structural diagram of an electronic device according to some embodiments of the present disclosure. With reference to Fig 3, in terms of hardware, the device includes a processor 302, an internal bus 304, a network interface 306, a memory 308, and a

non-transitory storage memory 310. Certainly, the device may further include hardware required by other functions. The processor 302 reads a corresponding computer program from the non-transitory storage memory 310 into the memory 308, and runs the computer program, so as to logically form a service providing apparatus based on a pre-training service system. Certainly, except for software implementations, at least one embodiment of the present disclosure does not exclude other embodiments, such as logic devices and combinations of software and hardware. In other words, an execution subject of the following processing flow is not limited to each logic unit, and may also be hardware or logic components.

Corresponding to the above method embodiment, as shown in FIG. 4, some embodiments of the present disclosure further provide a service providing apparatus based on a pre-training service system, where the pre-training service system includes a producer service module, an optimizer service module, and a consumer service module; a model maintained by the pre-training service system includes obfuscated model structure codes and obfuscated model parameters, and the obfuscated model structure codes and the obfuscated model parameters are generated after an obfuscation processing is pre-performed on original model structure codes and original model parameters respectively; and the apparatus includes:

an acquisition unit 410 arranged for acquiring identity information of a target user accessing the pre-training service system, so as to determine a preset user type to which the target user belongs; and

an opening unit 420 arranged for opening a service module matching the preset user type to which the target user belongs to the target user; where

the producer service module is opened to the target user in response to the target user being a model producer, where the producer service module is arranged for providing a model pre-training process for a pre-training dataset for the model producer to generate a pre-trained model;

the optimizer service module is opened to the target user in response to the target user being a model optimizer, where the optimizer service module is arranged for optimizing the pre-trained model according to a fine-tuning dataset provided by the model optimizer to obtain an optimized model; and

the consumer service module is opened to the target user in response to the target user being a model consumer, where the consumer service module is arranged for providing a service interface for at least one of the pre-trained model and the optimized model to the model consumer, the pre-trained model or the optimized model being arranged for performing inference on data

provided by the model consumer to output a model prediction result.

Optionally, the model pre-training function is used for performing training, according to the pre-training dataset, a general pre-trained model; where the pre-trained model includes at least one of the general pre-trained model and a dedicated pre-trained model generated by the producer service module according to the general pre-trained model and the dedicated pre-trained model corresponds to a downstream task instructed by the model producer.

Optionally, the service providing apparatus based on a pre-training service system further includes:

a generation unit 430 arranged for generating a code development template corresponding to the dedicated pre-trained model; and

a provision and modification unit 440 arranged for providing the code development template to the model optimizer; and modifying model structure codes and/or model parameters of the dedicated pre-trained model according to a model modification instruction issued by the model optimizer.

Optionally, the service providing apparatus based on a pre-training service system further includes:

an offline deployment unit 450 arranged for deploying the pre-trained model or the optimized model in a device corresponding to the model consumer through the service interface, for performing offline inference on the data provided by the model consumer; or,

an online inference unit 460 arranged for acquiring the data provided by the model consumer through the service interface, and invoke the pre-trained model or the optimized model to perform online inference.

Optionally, the obfuscation processing includes that: forward obfuscation processing is performed on at least a part of codes in the original model structure codes, and reverse obfuscation processing is performed on model parameters, associated with the at least a part of codes, in the original model parameters.

Optionally, the obfuscated model parameters are stored in a storage space managed by the pre-training service system.

Optionally, the service providing apparatus based on a pre-training service system further includes:

a first identification unit 470 arranged for identifying whether the model optimizer has a permission to invoke the obfuscated model parameters corresponding to the pre-trained model; and combining the obfuscated model parameters with the obfuscated model structure codes to obtain the pre-trained model and provide the pre-trained model to the model optimizer in response to the model optimizer having the permission to invoke the obfuscated model parameters; and/or,

a second identification unit 480 arranged for identifying whether the model consumer has a permission to invoke the obfuscated model parameters corresponding to the pre-trained model or the optimized model; and combining the obfuscated model parameters with the obfuscated model structure codes, so as to obtain the pre-trained model or the optimized model and provide the pre-trained model or the optimized model to the model consumer in response to the model consumer having the permission to invoke the obfuscated model parameters.

Optionally, the pre-training service system is further arranged for maintaining independent model access codes and model training inference codes for a target pre-trained model; where the model structure codes and the model access codes are encrypted and maintained by the pre-training service system, and the model training inference codes are maintained in plaintext by the pre-training service system.

Reference may be made to implementation processes of corresponding steps in the above method for specific implementation processes of functions and actions of each unit in the above apparatus, which will not be repeated herein.

The present disclosure further provides a computer program product. The computer program product includes a computer program, where when the computer program runs, the computer program causes a computer to execute the method in the above embodiment.

The present disclosure further provides a computer-readable storage medium. The computer-readable storage medium stores a computer program, where when executed by a processor, the computer program implements the method in the above embodiment.

The system, apparatus, module, or unit set forth in the above embodiments may be specifically implemented by a computer chip or entity or a product having a function. A specific form of a computer, a typical implementation device, may be a personal computer, a laptop computer, a cellular phone, a camera phone, a smart phone, a personal digital assistant, a media player, a navigation device, an email transmitting and receiving device, a game console, a tablet computer, a wearable device, or combinations of several ones of these devices.

In a typical configuration, the computer includes at least one central processing unit (CPU), an

input/output interface, a network interface, and a memory.

The memory may include a non-permanent memory such as a random access memory (RAM), and/or a non-transitory memory such as a read-only memory (ROM) or a flash RAM in a computer-readable medium. The memory is an instance of the computer-readable medium.

The computer-readable medium includes permanent, non-permanent, removable, and non-removable media, which can store information through any method or technology. The information can be a computer-readable instruction, a data structure, a module of a program, etc. Instances of the computer storage medium includes, but is not limited to, a phase-change random access memory (PRAM), a static random access memory (SRAM), a dynamic random access memory (DRAM), a random access memory (RAM) in another type, a read-only memory (ROM), an electrically erasable programmable read-only memory (EEPROM), a flash memory or another memory technology, a compact disc read-only memory (CD-ROM), a digital versatile disc (DVD) or another optical storage, a magnetic cartridge tape, a magnetic disk storage, a quantum memory, a graphene-based storage medium or other magnetic storage devices, or any other non-transmission media, and can be arranged for storing information that can be accessed by a calculation device. As defined herein, the computer-readable medium does not include transitory media, such as a modulated data signal and carrier wave.

In at least one embodiment of the present disclosure, the terms “comprise”, “include”, “encompass”, or their any other variations are intended to cover non-exclusive inclusions. Therefore, a process, method, article, or device including a series of elements further includes other elements that are not explicitly listed except for those elements, or further includes elements inherent to such a process, method, article, or device. Without more restrictions, the elements defined by the phrase “comprise a ...” and “include a ...” do not exclude the existence of other identical elements in the process, method, article, or device including the elements.

The terms used in at least one embodiment of the present disclosure are used for describing the particular embodiments, instead of limiting at least one embodiment of the present disclosure. The singular forms “a”, “an”, “the”, and “this” used in at least one embodiment of the present disclosure and the appended claims are also intended to include the plural forms, unless clearly stated in the context otherwise. It should also be understood that the term “and/or” used herein means and encompasses one or any or all possible combinations of multiple associated items listed.

It should be understood that while the terms first, second, third, etc. may be employed in at least one embodiment of the present disclosure to describe various information, these information

should not be limited thereto. These terms are used for distinguishing between the same type of information. For example, first information can also be referred to as second information, and similarly, second information can also be referred to as first information, without departing from the scope of at least one embodiment of the present disclosure. Depending on the context, the word “if” as used herein can be interpreted as “in a case of”, “when”, or “in response to determining”.

What are described above is at least one exemplary embodiment of the present disclosure, and are not intended to limit at least one embodiment the present disclosure. Any modifications, equivalent replacements, improvements, etc. made within the spirit and principles of at least one embodiment of the present disclosure should fall within the scope of protection of at least one embodiment the present disclosure.

Claims

1. A pre-training service system, wherein a model maintained by the pre-training service system comprises obfuscated model structure codes and obfuscated model parameters, and the obfuscated model structure codes and the obfuscated model parameters are generated after an obfuscation processing is pre-performed on original model structure codes and original model parameters respectively; and the pre-training service system comprises:

a producer service module arranged for providing a model pre-training process for a pre-training dataset for a model producer to generate a pre-trained model;

an optimizer service module arranged for optimizing the pre-trained model according to a fine-tuning dataset provided by a model optimizer to obtain an optimized model; and

a consumer service module arranged for providing a service interface for at least one of the pre-trained model and the optimized model for a model consumer, the pre-trained model or the optimized model being arranged for performing inference on data provided by the model consumer to output a model prediction result.

2. The system as claimed in claim 1, wherein the model pre-training function is used for performing training, according to the pre-training dataset, to obtain a general pre-trained model, wherein the pre-trained model comprises at least one of the general pre-trained model and a dedicated pre-trained model generated by the producer service module according to the general pre-trained model and the dedicated pre-trained model corresponds to a downstream task instructed by the model producer.

3. The system as claimed in claim 2, wherein the producer service module is further arranged for generating a code development template corresponding to the dedicated pre-trained model; and

the optimizer service module is further arranged for providing the code development template to the model optimizer; and modifying model structure codes and/or model parameters of the dedicated pre-trained model according to a model modification instruction issued by the model optimizer.

4. The system as claimed in any one of claims 1-3, wherein the consumer service module is

arranged for:

deploying the pre-trained model or the optimized model in a device corresponding to the model consumer through the service interface, for performing offline inference on data provided by the model consumer; or,

acquiring data provided by the model consumer through the service interface, and invoking the pre-trained model or the optimized model to perform online inference.

5. The system as claimed in any one of claims 1-4, wherein the obfuscation processing comprises:

performing forward obfuscation processing on at least a part of codes in the original model structure codes, and performing reverse obfuscation processing on model parameters, associated with the at least a part of codes, in the original model parameters.

6. The system as claimed in any one of claims 1-5, wherein the obfuscated model parameters are stored in a storage space managed by the pre-training service system.

7. The system as claimed in any one of claims 1-6, wherein

the optimizer service module is further arranged for identifying whether the model optimizer has a permission to invoke the obfuscated model parameters corresponding to the pre-trained model; and combining the obfuscated model parameters with the obfuscated model structure codes to obtain the pre-trained model and provide the pre-trained model to the model optimizer in response to the model optimizer having the permission to invoke the obfuscated model parameters; and/or,

the consumer service module is further arranged for identifying whether the model consumer has a permission to invoke the obfuscated model parameters corresponding to the pre-trained model or the optimized model; and combining the obfuscated model parameters with the obfuscated model structure codes to obtain the pre-trained model or the optimized model and provide the pre-trained model or the optimized model to the model consumer in response to the model consumer having the permission to invoke the obfuscated model parameters.

8. The system as claimed in any one of claims 1-7, wherein the pre-training service system is further arranged for maintaining independent model access codes and model training inference codes for the pre-trained model or the optimized model; wherein the model structure codes and the model access codes are encrypted and maintained by the pre-training service system, and the model training inference codes are maintained in plaintext by the pre-training service system.

9. A service providing method based on a pre-training service system, wherein the pre-training service system comprises a producer service module, an optimizer service module, and a consumer service module; a model maintained by the pre-training service system comprises obfuscated model structure codes and obfuscated model parameters, and the obfuscated model structure codes and the obfuscated model parameters are generated after an obfuscation processing is pre-performed on original model structure codes and original model parameters respectively; and the method comprises:

acquiring identity information of a target user accessing the pre-training service system to determine a preset user type to which the target user belongs; and

opening a service module matching the preset user type, to which the target user belongs, to the target user; wherein

the producer service module is opened to the target user in response to the target user being a model producer, wherein the producer service module is arranged for providing a model pre-training process for a pre-training dataset for the model producer to generate a pre-trained model;

the optimizer service module is opened to the target user in response to the target user being a model optimizer, wherein the optimizer service module is arranged for optimizing the pre-trained model according to a fine-tuning dataset provided by the model optimizer to obtain an optimized model; and

the consumer service module is opened to the target user in response to the target user being a model consumer, wherein the consumer service module is arranged for providing a service interface for at least one of the pre-trained model and the optimized model to the model consumer, the pre-trained model or the optimized model being arranged for performing inference on data provided by the model consumer to output a model prediction result.

10. The method as claimed in claim 9, wherein the model pre-training function is used for

performing training, according to the pre-training dataset, to obtain a general pre-trained model, wherein the pre-trained model comprises at least one of the general pre-trained model and a dedicated pre-trained model generated by the producer service module according to the general pre-trained model and the dedicated pre-trained model corresponds to a downstream task instructed by the model producer.

11. The method as claimed in claim 10, further comprising:

generating a code development template corresponding to the dedicated pre-trained model;

providing the code development template to the model optimizer; and modifying model structure codes and/or model parameters of the dedicated pre-trained model according to a model modification instruction issued by the model optimizer.

12. The method as claimed in any one of claims 9-11, further comprising:

deploying the pre-trained model or the optimized model in a device corresponding to the model consumer through the service interface, for performing offline inference on the data provided by the model consumer; or,

acquiring the data provided by the model consumer through the service interface, and invoking the pre-trained model or the optimized model to perform online inference.

13. The method as claimed in any one of claims 9-12, wherein the obfuscation processing comprises:

performing forward obfuscation processing on at least a part of codes in the original model structure codes, and performing reverse obfuscation processing on model parameters, associated with the at least a part of codes, in the original model parameters.

14. The method as claimed in any one of claims 9-13, wherein the obfuscated model parameters are stored in a storage space managed by the pre-training service system.

15. The method as claimed in any one of claims 9-14, further comprising:

identifying whether the model optimizer has a permission to invoke the obfuscated model parameters corresponding to the pre-trained model; and combining the obfuscated model parameters with the obfuscated model structure codes to obtain the pre-trained model and provide the pre-trained model to the model optimizer in response to the model optimizer having the permission to invoke the obfuscated model parameters; and/or,

identifying whether the model consumer has a permission to invoke the obfuscated model parameters corresponding to the pre-trained model or the optimized model; and combining the obfuscated model parameters with the obfuscated model structure codes to obtain the pre-trained model or the optimized model and provide the pre-trained model or the optimized model to the model consumer in response to the model consumer having the permission to invoke the obfuscated model parameters.

16. The method as claimed in any one of claims 9-15, wherein the pre-training service system is further arranged for maintaining independent model access codes and model training inference codes for a target pre-trained model; wherein the model structure codes and the model access codes are encrypted and maintained by the pre-training service system, and the model training inference codes are maintained in plaintext by the pre-training service system.

17. A service providing apparatus based on a pre-training service system, wherein the pre-training service system comprises a producer service module, an optimizer service module, and a consumer service module; a model maintained by the pre-training service system comprises obfuscated model structure codes and obfuscated model parameters, and the obfuscated model structure codes and the obfuscated model parameters are generated after an obfuscation processing is pre-performed on original model structure codes and original model parameters respectively; and the apparatus comprises:

an acquisition unit arranged for acquiring identity information of a target user accessing the pre-training service system, so as to determine a preset user type to which the target user belongs; and

an opening unit arranged for opening a service module matching the preset user type to which the target user belongs to the target user; wherein

the producer service module is opened to the target user in response to the target user being a model producer, wherein the producer service module is arranged for providing a model pre-training process for a pre-training dataset for the model producer to generate a pre-trained model;

the optimizer service module is opened to the target user in response to the target user being a model optimizer, wherein the optimizer service module is arranged for optimizing the pre-trained model according to a fine-tuning dataset provided by the model optimizer to obtain an optimized model; and

the consumer service module is opened to the target user in response to the target user being a model consumer, wherein the consumer service module is arranged for providing a service interface for at least one of the pre-trained model and the optimized model to the model consumer, the pre-trained model or the optimized model being arranged for performing inference on data provided by the model consumer to output a model prediction result.

18. An electronic device, comprising: a processor, a memory, and a non-transitory storage memory; wherein the processor is arranged for reading a computer program from the non-transitory storage memory into the memory, and executing the method as claimed in any one of claims 9-16 based on the computer program.

19. A computer-readable storage medium, storing a computer program, wherein when executed by a processor, the computer program implements the method as claimed in any one of claims 9-16.

20. A computer program product, comprising a computer program, wherein when the computer program runs, the computer program causes a computer to execute the method as claimed in any one of claims 9-16.