

VISUALIZATION METHODS AND RELATED DEVICES

CROSS-REFERENCE TO RELATED APPLICATION(S)

[1] This application claims priority to a Chinese Patent Application No. 202010656553.8 filed on July 09, 2020, and entitled “VISUALIZATION METHODS AND RELATED DEVICES”, the entire disclosure of which is hereby incorporated by reference.

TECHNICAL FIELD

[2] The present disclosure relates to the field of computer technology, and more particularly, to visualization methods and related devices.

BACKGROUND

[3] In recent years, a deep learning model has been widely used in image and video processing field. Generally, the deep learning model is obtained through training. However, the training process of the deep learning model involves complex and esoteric calculations and requires multiple iterations for implementation, which takes a long training time. In addition, performance of the deep learning model obtained through training can only be evaluated by developers at the end of the training.

SUMMARY

[4] Implementations of the present disclosure provide visualization methods and related devices to visualize training data of a deep learning model.

[5] According to a first aspect, a visualization method is provided in implementations of the present disclosure. The method is performed by a user equipment (UE) and includes the following. A visualization request for visualizing target training data of a target deep learning model is transmitted to a cloud server. A visualization response carrying visualized information of the target training data is received from the cloud server. The visualized information is displayed.

[6] According to a second aspect, a visualization apparatus is provided in implementations of the present disclosure. The apparatus is implemented in a UE and includes a transmitting unit, a receiving unit, and a display unit. The transmitting unit is configured to transmit a visualization request for visualizing target training data of a target deep learning model to a cloud server. The receiving unit is configured to receive a

visualization response carrying visualized information of the target training data from the cloud server. The display unit is configured to display the visualized information.

[7] According to a third aspect, a visualization method is provided in implementations of the present disclosure. The method is performed by a cloud server and includes the following. A visualization request for visualizing target training data of a target deep learning model is received from a UE. The target training data corresponding to the visualization request is obtained. Visualized information is obtained by pre-processing the target training data. A visualization response carrying the visualized information is transmitted to the UE.

[8] According to a fourth aspect, a visualization apparatus is provided in implementations of the present disclosure. The apparatus is implemented in a cloud server and includes a receiving unit, an obtaining unit, a determining unit, and a transmitting unit. The receiving unit configured to receive a visualization request for visualizing target training data of a target deep learning model from a UE. The obtaining unit is configured to obtain the target training data corresponding to the visualization request. The determining unit is configured to obtain visualized information by pre-processing the target training data. The transmitting unit is configured to transmit a visualization response carrying the visualized information to the UE.

[9] According to a fifth aspect, a computer device is provided in implementations of the present disclosure. The computer device includes a processor, a memory configured to store one or more programs, and a communication interface. The one or more programs are configured to be executed by the processor and include instructions configured to execute the operations of the methods described in the first aspect or the third aspect of the implementations of the present disclosure.

[10] According to a sixth aspect, a computer readable storage medium is provided in implementations of the present disclosure. The computer readable storage medium is configured to store computer programs for electrical data interexchange. The computer programs are operable with a computer to execute some or all of the operations of the methods described in the first aspect or the third aspect of the implementations of the present disclosure.

[11] According to a seventh aspect, a computer program product is provided in implementations of the present disclosure. The computer program product includes a non-transitory computer-readable storage medium that stores computer programs. The computer programs are operable with a computer to execute some or all of the operations of the methods described in the first aspect or the third aspect of the implementations of the present

disclosure. The computer program product may be a software installation package.

[12] It can be seen that, in implementations of the present disclosure, the UE transmits the visualization request to the cloud server, receives the visualization response carrying the visualized information of the target training data from the cloud server, and displays the visualized information. A training state of the deep learning model can be intuitively understood with the visualized information of the target training data, which helps to judge a feasibility of a current training strategy in time, thereby providing a basis for decision such as early stopping.

BRIEF DESCRIPTION OF THE DRAWINGS

[13] In order to describe technical solutions of implementations of the present disclosure or the related art more clearly, the following will give a brief introduction to the accompanying drawings required for describing implementations or the related art. Apparently, the accompanying drawings described hereinafter are some implementations of the present disclosure. Based on these drawings, those of ordinary skill in the art can also obtain other drawings without creative effort.

[14] FIG. 1 shows a schematic architecture diagram of a visualization system provided in implementations of the present disclosure.

[15] FIG. 2 shows a schematic flowchart of a visualization method provided in implementations of the present disclosure.

[16] FIG. 3 shows a schematic structural diagram of a computer device provided in implementations of the present disclosure.

[17] FIG. 4 shows a schematic structural diagram of a visualization apparatus provided in implementations of the present disclosure.

[18] FIG. 5 shows a schematic structural diagram of a visualization apparatus provided in other implementations of the present disclosure.

DETAILED DESCRIPTION

[19] In order to enable those skilled in the art to better understand solutions of the present disclosure, technical solutions in implementations of the present disclosure will be described clearly and completely hereinafter with reference to the accompanying drawings in the implementations of the present disclosure. Apparently, the described implementations are merely some rather than all implementations of the present disclosure. All other implementations obtained by those of ordinary skill in the art based on the implementations

of the present disclosure without creative efforts shall fall within the protection scope of the present disclosure.

[20] The following are described in detail.

[21] The terms “first”, “second”, “third”, “fourth”, and the like used in the specification, the claims, and the accompany drawings of the present disclosure are used to distinguish different objects rather than describe a particular order. The terms “include”, “comprise”, and “have” as well as variations thereof are intended to cover non-exclusive inclusion. For example, a process, method, system, product, or apparatus including a series of steps or modules is not limited to the listed steps or modules, on the contrary, it can optionally include other steps or modules that are not listed; alternatively, other steps or modules inherent to the process, method, product, or device can be included either.

[22] The term “implementation” referred to herein means that particular features, structures, or characteristics described in combination with implementations may be included in at least one implementation of the present disclosure. The phrase appearing in various places in the specification does not necessarily refer to the same implementation, nor does it refer to an independent or alternative implementation that is mutually exclusive with other implementations. It is expressly and implicitly understood by those skilled in the art that an implementation described herein may be combined with other implementations.

[23] Referring to FIG. 1, FIG. 1 shows a schematic architecture diagram of a visualization system provided in implementations of the present disclosure. The visualization system includes a supercomputing cluster, a cloud server, a cloud storage, and a user equipment (UE). It should be noted that, the form and the number of the supercomputing cluster, the cloud server, the cloud storage, and the UE shown in FIG. 1 are only used for examples and do not constitute a limitation to implementations of the present disclosure. Based on javascript and scalable vector graphics (svg) technology, the visualization system can visualize multi-dimensional and multi-class intermediate training data and support fast and efficient visualization rendering of a large amount of intermediate training data.

[24] The visualization system can provide production-level stable visualization system deployment and dynamic and controllable service capabilities based on a kubernetes container orchestration system. The kubernetes container is used to manage containerized applications on multiple hosts in the cloud server.

[25] The cloud server provides a visualization service. The visualization service is used to perform a data operation such as pre-processing on the obtained intermediate training data. The visualization rendering is performed on the intermediate training data on which the data

operation has been performed, to obtain visualized information.

[26] Optionally, the cloud server can also provide at least one of the following services: a service gateway, a training management service, a data storage service, and a permission service.

[27] The service gateway can serve as an entrance to at least one of the following services: a training management service, a data storage service, a data visualization service, and a permission service.

[28] Optionally, the service gateway may be an application program.

[29] Optionally, the service gateway can have a flow-limiting function.

[30] The training management service can provide at least one of the following services: monitoring whether a training task of a deep learning model is completed; recording, indexing, and searching an abnormal training task, comparing training processes of multiple training tasks of a same deep learning model, and sharing a training result of a deep learning model.

[31] The data storage service can provide at least one of the following services: data persisting, storing intermediate training data generated during deep learning model training by using a data thread, and storing the intermediate training data generated during the deep learning model training by caching.

[32] The permission management service is used to ensure security of the visualization system, and the visualization system can be used only after passing authentication of the permission management service. In addition, shared training is also authenticated via the permission management service which supports unified authentication account login and/or ordinary registration.

[33] The cloud server can interact with a developer toolkit via a hypertext transfer protocol (HTTP), a google remote procedure call (GRPC) protocol, or other protocols.

[34] The cloud server can directly store training data in the cloud storage.

[35] By running the developer toolkit, the supercomputing cluster can provide a development tool, for use together with a tensorflow, pytorch, caffe framework, and the like by a user, where the development tool includes a toolkit in a programming language, such as python. Additionally/alternatively, the supercomputing cluster can start a new thread as a data upload dedicated thread. Then the supercomputing cluster can upload the intermediate training data generated during the deep learning model training to the cloud server on the data upload dedicated thread via the HTTP communication protocol and the GRPC protocol. The uploaded intermediate training data may be a vector, a scalar, an image, a video, an audio,

and the like, and is stored in data cache of the developer toolkit.

[36] The computing speed of the deep learning model can be ensured by starting the new thread as the data upload dedicated thread.

[37] Optionally, the cloud storage may be a distributed storage and used for unified management of the intermediate training data generated by the deep learning model, and can provide massive storage for the intermediate training data. With the increase of a data volume of the intermediate training data, the cloud storage can dynamically expand its storage capacity to meet a storage demand of the intermediate training data, thereby providing subsequent guarantee for the increase of a visualization service.

[38] Optionally, the cloud storage may also be a data block-level cloud storage, a file-level cloud storage, an object-level cloud storage, and/or other forms of cloud storages.

[39] The UE can communicate with the cloud server via a preset application or a preset network address.

[40] Optionally, the UE may include various handheld devices, in-vehicle devices, wearable devices, and computing devices with a communication function or other processing devices connected to a wireless modem, as well as various forms of UEs, mobile stations (MS), terminal devices, and the like.

[41] Referring to FIG. 2, FIG. 2 shows a schematic flowchart of a visualization method provided in implementations of the present disclosure. The visualization method is performed by the UE and the cloud server above, and includes the following.

[42] At S201, the UE transmits a visualization request for visualizing target training data of a target deep learning model to the cloud server.

[43] At S202, the cloud server receives the visualization request for visualizing the target training data of the target deep learning model from the UE.

[44] In a possible implementation, the visualization request carries a training task identification of the target deep learning model. After receiving the visualization request, the cloud server can obtain related training data of the training task based on the training task identification of the target deep learning model. As an example, stored related training data of a training task of each deep learning model can be associated with a training task identification of the target deep learning model. For example, the training data can be stored according to the training task, and related training data of different training tasks can be stored at different positions. As another example, the training data can be stored according to its deep learning model type or data type.

[45] The data type may be a two-dimensional (2D) heat map, a three-dimensional (3D)

histogram, a feature map, an invoking sequence diagram, a scalar line graph, a directed acyclic graph (DAG), etc.

[46] Different data types correspond to different visualized icons.

[47] The training tasks are above the supercomputing cluster. The deep learning model training can be accelerated with a central processing unit of the supercomputing cluster.

[48] In this case, optionally, training data and a task identification of a training task corresponding to the training data can be stored in association, which will not be limited herein.

[49] In a possible implementation, the visualization request carries a training task identification of the target deep learning model and index information of the target training data. The index information is used to search the target training data, to determine a storage location of the target training data. The index information may be implemented in various methods. For example, the index information may include any one or more of: data creation time, a data type, a data label, etc. For another example, the index information includes a data type and a data label.

[50] There may be multiple data labels for a same data type, and the data label may be self-defined.

[51] For example, if the data type is a scalar line graph, the data label may be an exact value, a loss value, etc.

[52] The visualization request may be a request for sampling all data in an intermedia data set corresponding to the index information or randomly sampling data in the intermedia data set according to the training task identification and the index information.

[53] The target deep learning model includes at least one deep neural network. The target training data may be intermediate data generated by one deep neural network in the target deep learning model, or intermediate data generated by multiple deep neural networks in the target deep learning model.

[54] Optionally, the visualization request may be a request for visualizing intermediate data generated by the target deep learning model within a first time period. The first time period may have a duration of 3 min, 5 min, 9 min, 15 min, or the like, and an end moment of the first time period may be the current moment, earlier than the current moment, or later than the current moment.

[55] The target training data is stored in the cloud storage.

[56] In a possible implementation, the target training data includes at least one of: model performance variation trend information, model loss trend information, model

parameter distribution information, an intermediate result of model processing, model structure information, current progress information of model training, comparison information of different trainings for a same model, or a timing of scheduling time information of various processes.

[57] In a possible implementation, when the target training data includes at least one of: model performance variation trend information, model loss trend information, current progress information of model training, or comparison information of different trainings for a same model, the visualized information is a scalar line graph.

[58] When the target training data is model performance variation trend information, based on the scalar line graph, it can be determined whether the performance variation trend satisfies a first preset condition. If the performance variation trend satisfies the first preset condition, the model training is stopped. If the performance variation trend fails to satisfy the first preset condition, the model training is continued.

[59] When the target training data is model loss trend information, based on the scalar line graph, it can be determined whether the loss trend satisfies a second preset condition. If the loss trend satisfies the second preset condition, the model training is stopped. If the loss trend fails to satisfy the second preset condition, the model training is continued.

[60] When the target training data is current progress information of model training, based on the scalar line graph, it can be determined whether the training progress satisfies a preset progress. If the training progress satisfies the preset progress, the model training is continued. If the training progress fails to satisfy the preset progress, the model training is stopped.

[61] When the target training data is comparison information of different trainings for a same model, based on the scalar line graph, it can be determined whether features extracted from different trainings are the same. If the features are the same, the model training is continued. If the features are different, the model training is stopped.

[62] When the target training data is model parameter distribution information, the visualized information is a 3D histogram. Based on the 3D histogram, it can be determined whether the model parameter distribution is abnormal. If the model parameter distribution is abnormal, the model training is stopped.

[63] When the target training data is an intermediate result of model processing, the visualized information is a feature map. Based on the feature map, it can be determined whether a feature extracted by the model is correct. If the extracted feature is wrong, the model is reconstructed.

[64] When the target training data is model structure information, the visualized information is a DAG. Based on the DAG, it can be determined whether the structure of the model is correct. If the structure of the model is wrong, the structure of the model is reconstructed.

[65] An architecture of the deep learning model supported by the DAG may be a standard architecture such as open neural network exchange (ONNX), or other types of architectures. The ONNX may not need to perform data processing on the intermediate training data to obtain the DAG, while some architectures may need to perform pre-processing on the intermediate training data to obtain the DAG.

[66] When the target training data is scheduling time information of various processes, the visualized information is an invoking sequence diagram for various processes. Based on the invoking sequence diagram, time consumption of each operator can be determined. If the time consumption of the operator is longer than a preset time, the operator whose time consumption is longer than the preset time will be optimized.

[67] When the target training data is an intermediate result of model processing, the visualized information is a 2D heat map. Based on the 2D heat map, it can be determined whether a feature extracted by the model is correct. If the extracted feature is wrong, the model is reconstructed.

[68] The feature map and the 2D heat map are different forms of the target training data as the intermediate result of model processing.

[69] The type of the target training data includes at least one of: a scalar, a vector, an image, a video, or an audio.

[70] At S203, the cloud server obtains the target training data corresponding to the visualization request.

[71] In a possible implementation, the target training data corresponding to the visualization request is obtained as follows. Storage location information of the target training data is obtained by searching an index database based on the training task identification and the index information carried in the visualization request. The target training data stored in the cloud storage is obtained based on the storage location information of the target training data.

[72] A list of the intermediate training data is obtained based on the training task identification and the index information. The intermediate training data is obtained from the cloud storage based on the list of the intermediate training data.

[73] Optionally, the visualization request carries identification information of the target

training data of the target deep learning model. The identification information is used for the cloud server to obtain the target training data from the cloud storage. The identification information is preset and unique in the cloud storage.

[74] For example, the target deep learning model has training tasks including training task A and training task B. Training task A generates training data including training data A1 and training data A2. Training task B generates training data including training data B1 and training data B2. Training data A1, training data A2, training data B1, and training data B2 each are stored in a cloud platform. Training data A1 has identification information of 1, training data A2 has identification information of 2, training data B1 has identification information of 3, and training data B2 has identification information of 4. If the identification information is 1, the target training data obtained from the cloud storage by the cloud server is training data A1. If the identification information is 2, the target training data obtained from the cloud storage by the cloud server is training data A2. If the identification information is 3, the target training data obtained from the cloud storage by the cloud server is training data B1. If the identification information is 4, the target training data obtained from the cloud storage by the cloud server is training data B2.

[75] At S204, the cloud server obtains visualized information by pre-processing the target training data.

[76] At S205, the cloud server transmits a visualization response carrying the visualized information to the UE.

[77] At S206, the UE receives the visualization response carrying the visualized information of the target training data from the cloud server.

[78] At S207, the UE displays the visualized information.

[79] In a possible implementation, the visualized information includes at least one of: a DAG, a 3D histogram, an invoking sequence diagram for various processes, a feature map, a 2D heat map, or a scalar line graph.

[80] Optionally, when the visualized information is a DAG, the visualized information includes topology information of at least a portion of the target deep learning model and resource occupation information of each of at least one operator of the target deep learning model. The at least a portion of the target deep learning model includes multiple modules and/or multiple operators of the target deep learning model.

[81] Optionally, topology information of the multiple modules includes at least one of: identification information of the multiple modules, a dependency between the multiple modules, a data volume of each of the multiple modules, or information of at least one

operator of each of the multiple modules.

[82] The identification information of the module is unique and may be preset.

[83] Optionally, information of the operator includes at least one of: identification information of the operator, a dependency between the operators, or a data volume of the operator.

[84] The dependency may be a sequential relationship or a parallel relationship.

[85] Optionally, the resource occupation information is determined by the cloud server based on at least one of: a data type of the operator, input data information of the operator, or output data information of the operator.

[86] The input of the operator may be an image, a video, an audio, a scalar, a vector, etc. Different inputs correspond to different dimensions, for example, an image is equivalent to a 2D matrix, and an audio is equivalent to a one-dimensional matrix.

[87] The dimension of the input information of the operator may be changed after calculation by the operator.

[88] The data type of the operator may be a double precision type and may also be a single precision type.

[89] The operator may be at least one of: convolution, BatchNorm (BN), full connection, pooling, matrix multiplication and division, DropOut, activation, etc.

[90] The input data information may be an input data volume, and the output data information may be an output data volume.

[91] Optionally, when the visualized information is a scalar line graph, the visualized information includes at least one of: performance variation trend information of the target deep model, loss trend information of the target deep model, training progress information of the target deep learning model, or comparison information of different trainings in the target deep learning model.

[92] Optionally, when the visualized information is a feature map or a 2D heat map, the visualized information includes a feature extracted by the target deep learning model.

[93] Optionally, when the visualized information is a 3D histogram, the visualized information includes a parameter distribution of each of the multiple operators.

[94] Optionally, when the visualized information is an invoking sequence diagram for various processes, the visualized information includes a running duration of each of the multiple operators in each of at least one process.

[95] Optionally, the process may be at least one of: an interpretive language process, a local process, or an artificial intelligence (AI) chip process.

[96] The interpretive language process, the local process, and the AI chip process alternatively execute the invoking sequence diagram according to time.

[97] It can be seen that, in implementations of the present disclosure, the UE transmits the visualization request to the cloud server, receives the visualization response carrying the visualized information of the target training data from the cloud server, and displays the visualized information. A training state of the deep learning model can be intuitively understood with the visualized information of the target training data, which helps to judge a feasibility of a current training strategy in time, thereby providing a basis for decision such as early stopping.

[98] In an implementation of the present disclosure, the visualized information is displayed as follows. The visualized information is displayed in a folded display mode or an unfolded display mode. In the folded display mode, a module of the target deep learning model serves as a minimum display unit, where the module includes at least one operator. Additionally/alternatively, in the unfolded display mode, an operator of the target deep learning model serves as a minimum display unit.

[99] When the visualized information is a DAG, the visualized information can be displayed in the folded display mode or the unfolded display mode. When the visualized information is not a DAG, the visualized information is displayed in the unfolded display mode.

[100] In the unfolded display mode, the module may display the operator(s) in the module after receiving an operation instruction.

[101] When the visualized information is a 3D histogram, in the unfolded display mode, a parameter distribution of each operator in each module may be directly displayed.

[102] When the visualized information is an invoking sequence diagram, in the unfolded display mode, a running duration of each operator in each module in at least one process may be directly displayed.

[103] For example, assume that the visualized information is a DAG, the target deep learning model includes two modules (i.e., module A and module B), module A includes three operators (i.e., A1, A2, and A3), and module B includes two operators (i.e., B1 and B2). Then, if the visualized information is displayed in the folded display mode, module A and module B are displayed. If the visualized information is displayed in the unfolded display mode, operators A1, A2, and A3 in module A are displayed after module A receives an operation instruction, and operators B1 and B2 in module B are displayed after module B receives an operation instruction.

[104] For example, assume that the visualized information is a 3D histogram, the target deep learning model includes two modules (i.e., module A and module B), module A includes three operators (i.e., A1, A2, and A3), module B includes two operators (i.e., B1 and B2), module A corresponds to 3D histogram 1, and module B corresponds to 3D histogram B. Then, 3D histogram 1 includes a parameter distribution of each of operators A1, A2, and A3. 3D histogram 2 includes a parameter distribution of each of operators B1 and B2.

[105] For example, assume that the visualized information is an invoking sequence diagram, the target deep learning model includes two modules (i.e., module A and module B) and two processes (i.e., processes C1 and C2), module A includes two operators (i.e., A1 and A2), module B includes two operators (i.e., B1 and B2), module A corresponds to invoking sequence diagram 1, and module B corresponds to an invoking sequence diagram B. Then, the invoking sequence diagram 1 includes a running duration of operator A1 in each of processes C1 and C2, and a running duration of operator A2 in each of processes C1 and C2. Invoking sequence diagram 2 includes a running duration of operator B1 in each of processes C1 and C2, and a running duration of operator B2 in each of processes C1 and C2.

[106] Optionally, the display mode of the visualized information can be preset. The visualized information can be displayed in the folded display mode, and then in the unfolded display mode after a first duration. The first duration can be preset.

[107] Optionally, when the visualized information is a scalar line graph, the line graph may have multiple lines which can be scaled or translated after an operation instruction is received. In the unfolded display mode, the scalar line graph may directly display at least one of: a performance variation trend of the target deep model, a loss trend of the target deep model, a training progress of the target deep learning model, or features extracted from different trainings in the target deep learning model.

[108] Optionally, when the visualized information is a feature map or a 2D heat map, in the unfolded display mode, a feature extracted by the target deep learning model may be directly displayed.

[109] In the 2D heat map, a higher heat represents a higher probability of being a feature.

[110] It can be seen that, in implementations of the present disclosure, the UE displays the visualized information in different methods, which is beneficial to improve an application range of the UE.

[111] In an implementation of the present disclosure, the method further includes the following. Before the visualization request is transmitted to the cloud server, the intermediate

training data of the target deep learning model is transmitted to the cloud server via a data upload thread, for the cloud server to obtain the index information of the target training data and a training task identification. The index information is stored into a database based on the training task identification. The target training data is stored into a cloud storage based on the index information.

[112] The data upload thread is a newly started thread.

[113] The training task identification is unique. The training task identification may be determined after the intermediate training data is generated, or may be determined at the start of the training.

[114] The intermediate training data includes the target training data, the training task identification, and the index information.

[115] Optionally, the intermediate training data of the target deep learning model is transmitted to the cloud server via the data upload thread, for the cloud server to obtain the identification information of the target training data. The identification information is stored into the database. The target training data is stored into the cloud storage based on the identification information.

[116] The identification information is unique.

[117] Optionally, the method further includes the following. Before the intermediate training data of the target deep learning model is transmitted to the cloud server via the data upload thread, a data volume of the target training data in the intermediate training data is determined. When the data volume is greater than a preset data volume, the index information of the target training data and the training task identification are determined, and the index information of the target training data and the training task information are carried on the intermediate training data. When the data volume is less than the preset data volume, the identification information of the target training data is determined, and the identification information is carried on the intermediate training data.

[118] Optionally, when the cloud server receives the identification information, the identification information can be stored into a first buffer of the database, and the target training data can be stored into a second buffer of the database based on the identification information.

[119] It can be seen that, in implementations of the present disclosure, the target data is stored in the cloud storage via the cloud server, which is beneficial for the cloud server to quickly obtain the target training data after receiving the visualization request.

[120] In an implementation of the present disclosure, the visualized information is

obtained by pre-processing the target training data as follows. Pre-processed data is obtained by performing at least one pre-processing on the target training data. The visualized information is obtained by rendering the pre-processed data.

[121] If the data type of the target training data is an image or a video, the pre-processing on the image or the video includes one or more of: insensitive area rejection processing, image accuracy enhancement processing, image noise reduction processing, or image binarization processing.

[122] If the data type of the target training data is a text, the pre-processing on the text includes one or more of: document segmentation processing, text word segmentation processing, stop word removal processing, where the stop word includes a punctuation, a number, a single character, and other meaningless words, text feature extraction, word frequency statistics processing, or text vectorization processing.

[123] The visualization rendering is to assemble the pre-processed data into a hypertext markup language (HTML).

[124] The visualized information obtained via the pre-processing is not a DAG.

[125] It can be seen that, in implementations of the present disclosure, the pre-processed data is obtained by pre-processing the target training data, and the visualized information is obtained by rendering the pre-processed data, which is beneficial to analyze the target deep learning model based on the visualized information.

[126] In an implementation of the present disclosure, the visualized information is obtained by pre-processing the target training data as follows. The visualized information is obtained by analyzing via a binary tree the target training data on condition that a training task of the target deep learning model is executed in a target framework.

[127] The target deep learning model processing the target training data via the binary tree is of type parrots.

[128] The visualized information obtained by analyzing via the binary tree the target training data is a DAG.

[129] It can be seen that, in implementations of the present disclosure, the visualized information is obtained by analyzing the target training data, which is beneficial to analyze the target deep learning model based on the visualized information.

[130] Referring to FIG.3, FIG. 3 shows a schematic structural diagram of a computer device provided in implementations of the present disclosure. The computer device includes a processor, a memory configured to store one or more programs, and a communication interface. The one or more programs are configured to be executed by the processor.

[131] In an implementation, the computer device is a UE, and the programs include instructions configured to execute the following operations. A visualization request for visualizing target training data of a target deep learning model is transmitted to a cloud server. A visualization response carrying visualized information of the target training data is received from the cloud server. The visualized information is displayed.

[132] Optionally, the visualized information includes at least one of: a DAG, a 3D histogram, an invoking sequence diagram for various processes, a feature map, a 2D heat map, or a scalar line graph. Additionally/alternatively, the target training data includes at least one of: model performance variation trend information, model loss trend information, model parameter distribution information, an intermediate result of model processing, model structure information, current progress information of model training, comparison information of different trainings for a same model, or scheduling time information of various processes.

[133] Optionally, the visualized information includes at least one of: topology information of at least a portion of the target deep learning model, or resource occupation information of each of at least one operator of the target deep learning model. The at least a portion of the target deep learning model includes multiple modules and/or multiple operators of the target deep learning model.

[134] Optionally, topology information of the multiple modules includes at least one of: identification information of the multiple modules, a dependency between the multiple modules, a data volume of each of the multiple modules, or information of at least one operator of each of the multiple modules.

[135] Optionally, the resource occupation information is determined by the cloud server based on at least one of: a data type of the operator, input data information of the operator, or output data information of the operator.

[136] Optionally, the visualization request carries a training task identification of the target deep learning model and index information of the target training data, and the index information includes a data type and a data label.

[137] Optionally, in terms of displaying the visualized information, the programs include instructions configured to execute the following operations.

[138] In an implementation of the present disclosure, visualized data can be displayed in different modes. In some implementations, the visualized information can be displayed in a folded display mode or an unfolded display mode, or other display modes may also be defined, which will not be limited herein.

[139] In some implementations, in the folded display mode, a module of the target deep learning model serves as a minimum display unit, where the module includes at least one operator. The modules here can be divided by a default division method, or can also be set by a user, which will not be limited herein.

[140] In some implementations, in the unfolded display mode, an operator of the target deep learning model serves as a minimum display unit. In this case, information of all operators in the model can be displayed.

[141] In some implementations, the modules can also be displayed in a mixed display mode, i.e., some modules are displayed in the folded display mode and other modules are displayed in the unfolded display mode. Alternatively, the modules can be displayed based on user settings, which will not be limited herein.

[142] Optionally, the programs include instructions configured to execute the following operations. Before the visualization request is transmitted to the cloud server, intermediate training data of the target deep learning model is transmitted to the cloud server via a data upload thread, for the cloud server to obtain index information of the target training data and a training task identification. The index information is stored into a database based on the training task identification. The target training data is stored into a cloud storage based on the index information.

[143] In another implementation, the computer device is a cloud server, and the programs include instructions configured to execute the following operations. A visualization request for visualizing target training data of a target deep learning model is received from a UE. The target training data corresponding to the visualization request is obtained. Visualized information is obtained by pre-processing the target training data. A visualization response carrying the visualized information is transmitted to the UE.

[144] Optionally, the visualized information includes at least one of: a DAG, a 3D histogram, an invoking sequence diagram for various processes, a feature map, a 2D heat map, or a scalar line graph. Additionally/alternatively, the target training data includes at least one of: model performance variation trend information, model loss trend information, model parameter distribution information, an intermediate result of model processing, model structure information, current progress information of model training, comparison information of different trainings for a same model, or scheduling time information of various processes.

[145] Optionally, the visualized information includes at least one of: topology information of at least a portion of the target deep learning model, or resource occupation

information of each of at least one operator of the target deep learning model. The at least a portion of the target deep learning model includes multiple modules and/or multiple operators of the target deep learning model.

[146] Optionally, topology information of the multiple modules includes at least one of: identification information of the multiple modules, a dependency between the multiple modules, a data volume of each of the multiple modules, or information of at least one operator of each of the multiple modules.

[147] Optionally, in terms of obtaining the visualized information by pre-processing the target training data, the programs include instructions configured to execute the following operations. Resource occupation information of each of the at least one operator is determined based on at least one of: a data type of at least one operator of the target deep learning model, input data information of the operator, or output data information of the operator.

[148] Optionally, the visualization request carries a training task identification of the target deep learning model and index information of the target training data, and the index information includes a data type and a data label.

[149] In terms of obtaining the target training data corresponding to the visualization request, the programs include instructions configured to execute the following operations. Storage location information of the target training data is obtained by searching, based on the training task identification and the index information carried in the visualization request, an index database. Based on the storage location information of the target training data, the target training data stored in a cloud storage is obtained.

[150] Optionally, in terms of obtaining the visualized information by pre-processing the target training data, the programs include instructions configured to execute the following operations. Pre-processed data is obtained by performing at least one pre-processing on the target training data. The visualized information is obtained by rendering the pre-processed data.

[151] Optionally, in terms of obtaining the visualized information by pre-processing the target training data, the programs include instructions configured to execute the following operations. The visualized information is obtained by analyzing via a binary tree the target training data on condition that a training task of the target deep learning model is executed in a target framework.

[152] It is to be noted that, for the implementation process of these implementations, reference can be made to the implementation process in the above method implementations,

which will not be repeated herein.

[153] Referring to FIG. 4, FIG. 4 illustrates a visualization apparatus provided in implementations of the present disclosure. The visualization apparatus is implemented in a UE and includes a transmitting unit 401, a receiving unit 402, and a display unit 403. The transmitting unit 401 is configured to transmit a visualization request for visualizing target training data of a target deep learning model to a cloud server. The receiving unit 402 is configured to receive a visualization response carrying visualized information of the target training data from the cloud server. The display unit 403 is configured to display the visualized information.

[154] In an implementation of the present disclosure, the visualized information includes at least one of: a DAG, a 3D histogram, an invoking sequence diagram for various processes, a feature map, a 2D heat map, or a scalar line graph. Additionally/alternatively, the target training data includes at least one of: model performance variation trend information, model loss trend information, model parameter distribution information, an intermediate result of model processing, model structure information, current progress information of model training, comparison information of different trainings for a same model, or scheduling time information of various processes.

[155] In an implementation of the present disclosure, the visualized information includes at least one of: topology information of at least a portion of the target deep learning model, or resource occupation information of each of at least one operator of the target deep learning model. The at least a portion of the target deep learning model includes multiple modules and/or multiple operators of the target deep learning model.

[156] In an implementation of the present disclosure, topology information of the multiple modules includes at least one of: identification information of the multiple modules, a dependency between the multiple modules, a data volume of each of the multiple modules, or information of at least one operator of each of the multiple modules.

[157] In an implementation of the present disclosure, the resource occupation information is determined by the cloud server based on at least one of: a data type of the operator, input data information of the operator, or output data information of the operator.

[158] In an implementation of the present disclosure, the visualization request carries a training task identification of the target deep learning model and index information of the target training data, and the index information includes a data type and a data label.

[159] In an implementation of the present disclosure, in terms of displaying the visualized information, the display unit 403 is configured to display the visualized

information in a folded display mode or an unfolded display mode. In the folded display mode, a module of the target deep learning model serves as a minimum display unit, where the module includes at least one operator. Additionally/alternatively, in the unfolded display mode, an operator of the target deep learning model serves as a minimum display unit.

[160] In an implementation of the present disclosure, the transmitting unit 401 is further configured to transmit intermediate training data of the target deep learning model via a data upload thread to the cloud server before transmitting the visualization request to the cloud server. The intermediate training data is used for the cloud server to obtain index information of the target training data and a training task identification.

[161] The apparatus further includes a data storage unit 404. The data storage unit 404 is configured to store the index information into a database based on the training task identification and store the target training data into a cloud storage based on the index information.

[162] It is to be noted that, the transmitting unit 401 and the receiving unit 402 can be implemented by a communication interface. The display unit 403 can be implemented by a display screen. The data storage unit 404 can be implemented by a processor.

[163] Referring to FIG.5, FIG. 5 illustrates a visualization apparatus provided in implementations of the present disclosure. The visualization apparatus is implemented in a cloud server and includes a receiving unit 501, an obtaining unit 502, a determining unit 503, and a transmitting unit 504. The receiving unit 501 is configured to receive a visualization request for visualizing target training data of a target deep learning model from a UE. The obtaining unit 502 is configured to obtain the target training data corresponding to the visualization request. The determining unit 503 is configured to obtain visualized information by pre-processing the target training data. The transmitting unit 504 is configured to transmit a visualization response carrying the visualized information to the UE.

[164] In an implementation of the present disclosure, the visualized information includes at least one of: a DAG, a 3D histogram, an invoking sequence diagram for various processes, a feature map, a 2D heat map, or a scalar line graph. Additionally/alternatively, the target training data includes at least one of: model performance variation trend information, model loss trend information, model parameter distribution information, an intermediate result of model processing, model structure information, current progress information of model training, comparison information of different trainings for a same model, or scheduling time information of various processes.

[165] In an implementation of the present disclosure, the visualized information

includes at least one of: topology information of at least a portion of the target deep learning model, or resource occupation information of each of at least one operator of the target deep learning model. The at least a portion of the target deep learning model includes multiple modules and/or multiple operators of the target deep learning model.

[166] In an implementation of the present disclosure, topology information of the multiple modules includes at least one of: identification information of the multiple modules, a dependency between the multiple modules, a data volume of each of the multiple modules, or information of at least one operator of each of the multiple modules.

[167] In an implementation of the present disclosure, in terms of obtaining the visualized information by pre-processing the target training data, the programs include instructions configured to execute the following operations. Resource occupation information of each of the at least one operator is determined based on at least one of: a data type of at least one operator of the target deep learning model, input data information of the operator, or output data information of the operator.

[168] In an implementation of the present disclosure, the visualization request carries a training task identification of the target deep learning model and index information of the target training data, and the index information includes a data type and a data label. In terms of obtaining the target training data corresponding to the visualization request, the obtaining unit 502 is configured to obtain storage location information of the target training data by searching, based on the training task identification and the index information carried in the visualization request, an index database. The obtaining unit 502 is configured to obtain the target training data stored in a cloud storage based on the storage location information of the target training data.

[169] In an implementation of the present disclosure, in terms of obtaining the visualized information by pre-processing the target training data, the determining unit 503 is configured to obtain pre-processed data by performing at least one pre-processing on the target training data, and obtain the visualized information by rendering the pre-processed data.

[170] In an implementation of the present disclosure, in terms of obtaining the visualized information by pre-processing the target training data, the determining unit 503 is configured to obtain the visualized information by analyzing via a binary tree the target training data on condition that a training task of the target deep learning model is executed in a target framework.

[171] It is to be noted that, the receiving unit 501 and the transmitting unit 504 can be

implemented by a communication interface. The obtaining unit 502 and the determining unit 503 can be implemented by a processor.

[172] Implementations further provide a computer readable storage medium. The computer readable storage medium is configured to store computer programs configured to perform electronic data interchange. The computer programs are operable with a computer to execute some or all of the operations described by the electronic device or the cloud server in the foregoing method implementations.

[173] Implementations further provide a computer program product. The computer program product includes a non-transitory computer-readable storage medium that stores computer programs. The computer programs are operable with a computer to execute some or all of the operations described by the electronic device or the cloud server in the foregoing method implementations. The computer program product may be a software installation package.

[174] The algorithmic operations and the method described in implementations may be implemented as a hardware, or may be implemented by executing a software instruction with a processor. The software instruction may be composed with corresponding software modules. The software module can be stored in a storage medium such as a random access memory (RAM), a flash memory, a read only memory (ROM), an erasable programmable ROM (EPROM), an electrically EPROM (EEPROM), a register, a hard disk, a mobile hard disk, or a compact disc ROM (CD-ROM), or in storage media with any other forms known in this field. An exemplary storage medium is coupled to the processor, such that the processor can read information from the storage medium and write information to the storage medium. The storage medium can also be a component of the processor. The processor and the storage medium may be located in an application specific integrated circuit (ASIC). In addition, the ASIC can be located in an access network device, a target network device, or a core network device. The processor and the storage medium may also be present as discrete components in the access network device, the target network device, or the core network device.

[175] Those skilled in the art can be aware that, in the above mentioned one or more examples, all or part of the above implementations can be implemented through software, hardware, firmware, or any other combination thereof. When implemented by software, all or part of the above implementations can be implemented in the form of a computer program product. The computer program product includes one or more computer instructions. When the computer instructions are applied and executed on a computer, all or part of the operations or functions of the implementations of the present disclosure are performed. The

computer can be a general-purpose computer, a special-purpose computer, a computer network, or other programmable apparatuses. The computer instruction can be stored in a computer readable storage medium, or transmitted from one computer readable storage medium to another computer readable storage medium. For example, the computer instruction can be transmitted from one website, computer, server, or data center to another website, computer, server, or data center in a wired manner or in a wireless manner. Examples of the wired manner can be a coaxial cable, an optical fiber, a digital subscriber line (DSL), etc. The wireless manner can be, for example, infrared, wireless, microwave, etc. The computer readable storage medium can be any computer accessible usable-medium or a data storage device such as a server, a data center, or the like which is integrated with one or more usable media. The usable medium can be a magnetic medium (such as a soft disc, a hard disc, or a magnetic tape), an optical medium (such as a digital video disc (DVD)), or a semiconductor medium (such as a solid-state disk (SSD)), etc.

[176] The above-mentioned implementations provide further details of the purpose, technical solutions, and beneficial effects of the implementations of the present disclosure. It can be understood that, the above is only an implementation of the present disclosure and is not intended to limit the scope of protection of the present disclosure. Any modification, equivalent arrangements and improvement made within the spirit and principles of the present disclosure shall be included in the scope of protection of the present disclosure.

What is claimed is:

1. A visualization method performed by a user equipment (UE), comprising:
transmitting, to a cloud server, a visualization request for visualizing target training data of a target deep learning model;
receiving, from the cloud server, a visualization response carrying visualized information of the target training data; and
displaying the visualized information.
2. The method of claim 1, wherein
the visualized information comprises at least one of: a directed acyclic graph (DAG), a three-dimensional (3D) histogram, an invoking sequence diagram for various processes, a feature map, a two-dimensional (2D) heat map, or a scalar line graph; and/or
the target training data comprises at least one of: model performance variation trend information, model loss trend information, model parameter distribution information, an intermediate result of model processing, model structure information, current progress information of model training, comparison information of different trainings for a same model, or scheduling time information of various processes.
3. The method of claim 1 or 2, wherein the visualized information comprises at least one of:
topology information of at least a portion of the target deep learning model, wherein the at least a portion of the target deep learning model comprises a plurality of modules and/or a plurality of operators of the target deep learning model; or
resource occupation information of each of at least one operator of the target deep learning model.
4. The method of claim 3, wherein topology information of the plurality of modules comprises at least one of:
identification information of the plurality of modules, a dependency between the plurality of modules, a data volume of each of the plurality of modules, or information of at least one operator of each of the plurality of modules.
5. The method of claim 3 or 4, wherein the resource occupation information is

determined by the cloud server based on at least one of: a data type of the operator, input data information of the operator, or output data information of the operator.

6. The method of any of claims 1 to 5, wherein the visualization request carries a training task identification of the target deep learning model and index information of the target training data, and the index information comprises a data type and a data label.

7. The method of any of claims 1 to 6, wherein displaying the visualized information comprises:

displaying the visualized information in a folded display mode or an unfolded display mode, wherein

in the folded display mode, a module of the target deep learning model serves as a minimum display unit, wherein the module comprises at least one operator;
and/or

in the unfolded display mode, an operator of the target deep learning model serves as a minimum display unit.

8. The method of any of claims 1 to 7, further comprising:

before transmitting, to the cloud server, the visualization request,

transmitting, to the cloud server, intermediate training data of the target deep learning model via a data upload thread, for the cloud server to obtain index information of the target training data and a training task identification;

storing, based on the training task identification, the index information into an index database; and

storing, based on the index information, the target training data into a cloud storage.

9. A visualization method performed by a cloud server, comprising:

receiving, from a user equipment (UE), a visualization request for visualizing target training data of a target deep learning model;

obtaining the target training data corresponding to the visualization request;

obtaining visualized information by pre-processing the target training data; and

transmitting, to the UE, a visualization response carrying the visualized information.

10. The method of claim 9, wherein

the visualized information comprises at least one of: a directed acyclic graph (DAG), a three-dimensional (3D) histogram, an invoking sequence diagram for various processes, a feature map, a two-dimensional (2D) heat map, or a scalar line graph; and/or

the target training data comprises at least one of: model performance variation trend information, model loss trend information, model parameter distribution information, an intermediate result of model processing, model structure information, current progress information of model training, comparison information of different trainings for a same model, or scheduling time information of various processes.

11. The method of claim 9 or 10, wherein the visualized information comprises at least one of:

topology information of at least a portion of the target deep learning model, wherein the at least a portion of the target deep learning model comprises a plurality of modules and/or a plurality of operators of the target deep learning model; or

resource occupation information of each of at least one operator of the target deep learning model.

12. The method of claim 11, wherein topology information of the plurality of modules comprises at least one of:

identification information of the plurality of modules, a dependency between the plurality of modules, a data volume of each of the plurality of modules, or information of at least one operator of each of the plurality of modules.

13. The method of any of claims 9 to 12, wherein obtaining the visualized information by pre-processing the target training data comprises:

determining, based on at least one of: a data type of at least one operator of the target deep learning model, input data information of the operator, or output data information of the operator, resource occupation information of each of the at least one operator.

14. The method of any of claims 9 to 13, wherein

the visualization request carries a training task identification of the target deep learning model and index information of the target training data, and the index information

comprises a data type and a data label; and

obtaining the target training data corresponding to the visualization request comprises:

obtaining storage location information of the target training data by searching, based on the training task identification and the index information carried in the visualization request, an index database; and

obtaining, based on the storage location information of the target training data, the target training data stored in a cloud storage.

15. The method of any of claims 9 to 14, wherein obtaining the visualized information by pre-processing the target training data comprises:

obtaining pre-processed data by performing at least one pre-processing on the target training data; and

obtaining the visualized information by rendering the pre-processed data.

16. The method of any of claims 9 to 15, wherein obtaining the visualized information by pre-processing the target training data comprises:

obtaining the visualized information by analyzing via a binary tree the target training data on condition that a training task of the target deep learning model is executed in a target framework.

17. A visualization apparatus, implemented in a user equipment (UE) and comprising:

a transmitting unit configured to transmit, to a cloud server, a visualization request for visualizing target training data of a target deep learning model;

a receiving unit configured to receive, from the cloud server, a visualization response carrying visualized information of the target training data; and

a display unit configured to display the visualized information.

18. A visualization apparatus, implemented in a cloud server and comprising:

a receiving unit configured to receive, from a user equipment (UE), a visualization request for visualizing target training data of a target deep learning model;

an obtaining unit configured to obtain the target training data corresponding to the visualization request;

a determining unit configured to obtain visualized information by pre-processing the target training data; and

a transmitting unit configured to transmit, to the UE, a visualization response carrying the visualized information.

19. A computer device, comprising a processor, a memory configured to store one or more programs, and a communication interface, wherein the one or more programs are configured to be executed by the processor and comprise instructions configured to perform the method of any of claims 1-8 or 9-16.

20. A computer-readable storage medium storing computer programs, wherein when executed, the computer programs are configured to perform the method of any of claims 1-8 or 9-16.