COMMUNICATION METHOD, APPARATUS, AND SYSTEM

[0001] This application claims priority to Chinese Patent Application No. 202410332551.1, filed with the China National Intellectual Property Administration on March 21, 2024 and entitled "COMMUNICATION METHOD, APPARATUS, AND SYSTEM", which is incorporated herein by reference in its entirety.

TECHNICAL FIELD

[0002] This application relates to the field of communication technologies, and more specifically, to a communication method, apparatus, and system.

BACKGROUND

[0003] The rapid development of large models has led to the emergence of applications in various industries. A cloud-side large model may be a large model deployed on a cloud-side device or an edge-side device. A terminal-side large model may be a large model deployed on a terminal-side device. The terminal-side large model can meet user requirements for protecting terminal-side privacy data, thereby attracting widespread attention. However, due to limited storage space, the terminal-side device can only store a large model with a small parameter quantity, which performs much worse than a large model with a large parameter quantity, hindering the application of the large model on the terminal side.

[0004] Therefore, in an inference process of the terminal-side large model, how to cope with a contradiction between an excessively large parameter quantity of the large model and excessively small storage space, to improve inference performance of the terminal-side large model is an urgent problem to be resolved.

SUMMARY

[0005] This application provides a communication method, apparatus, and system, to improve inference performance of a terminal-side large model.

[0006] According to a first aspect, a communication method is provided. The method may be performed by a first apparatus (for example, a terminal device, or a terminal-side device), may be performed by a component (for example, a processor, a chip, or a system on a chip) in the first apparatus, or may be performed by a logical module or software that can implement all or a part

of functions of the first apparatus.

[0007] The method includes: receiving first information, where the first information indicates a first active parameter, the first active parameter is an active parameter of N first modules in a first model, and N is a positive integer; and performing inference based on the N first modules according to the first active parameter.

[0008] According to the foregoing embodiment, the first apparatus for inference does not need to store full parameters of the first module, but may obtain active parameters in the full parameters of the first module through the first information. Compared with a solution in which the full parameters are stored locally in the first apparatus and the active parameters are read from the full parameters, the foregoing embodiment can save storage space for the first apparatus, so that the first apparatus with limited storage space can run a large model with a large parameter quantity, thereby improving inference performance of the large model. For example, the first module may be a feedforward neural network (feedforward neural network, FNN) module in a large model of a transformer (transformer) structure. A ratio of full parameters of the FNN module to full parameters of a multi-head attention (multi-head attention, MHA) module is about 2:1. If the full parameters of the FNN module are not stored in the terminal-side device, two thirds of flash memory space of the terminal-side device is saved. The saved flash memory space may store a parameter of another module, for example, the parameter of the MHA module. In this way, a quantity of parameters of the large model can be increased by three times, thereby effectively improving performance of the large model.

[0009] In some implementations, the N first modules include N1 first modules, the N1 first modules are modules at a first layer of at least one layer in the first model, and N1 is a positive integer less than or equal to N.

[0010] According to the foregoing embodiment, the first apparatus may perform inference based on the module at the first layer in the first model according to the active parameter indicated by the first information. Compared with a solution in which inference is performed based on the full parameters and the module at the first layer, this embodiment of this application does not require the full parameters of the module at the first layer to be loaded into a memory, so that memory space of the first apparatus is saved, thereby further improving inference performance of the first model.

[0011] In some implementations, the N1 first modules are a 2nd one of a plurality of modules at the first layer in the first model, and N1=1. The method further includes: predicting information about a location of the first active parameter based on a second input value, where the second input value is an input value of the 2nd one of the plurality of modules at the first layer in the first model or an output value of a 1st one of the plurality of modules at the first layer in the first model.

[0012] According to the foregoing embodiment, the first apparatus may predict an active parameter of the 2nd one of the plurality of modules at the first layer in the first module, to perform inference based on the 2nd one of the plurality of modules according to the active parameter of the 2nd one of the plurality of modules. Compared with a solution in which inference is performed based on the full parameters and the 2nd one of the plurality of modules at the first layer, this embodiment of this application does not require the full parameters of the 2nd one of the plurality of modules at the first layer to be loaded into a memory, so that memory space of the first apparatus is saved, thereby further improving inference performance of the first model.

[0013] In some implementations, the method further includes: predicting information about a location of a second active parameter based on a first input value, where the second active parameter is an active parameter of the 1st one of the plurality of modules at the first layer in the first model, and the first input value is an input value of the 1st one of the plurality of modules at the first layer in the first model.

[0014] According to the foregoing embodiment, the first apparatus may predict the active parameter of the 1st one of the plurality of modules the first layer in the first module, to perform inference based on the 1st one of the plurality of modules according to the active parameter of the 1st one of the plurality of modules. Compared with a solution in which inference is performed based on the full parameters and the 1st one of the plurality of modules at the first layer, this embodiment of this application does not require the full parameters of the 1st one of the plurality of modules at the first layer to be loaded into a memory, so that memory space of the first apparatus is saved, thereby further improving inference performance of the first model.

[0015] In some implementations, the N first modules further include N2 first modules, the first model includes N2 layers, each of the N2 layers includes one of the N2 first modules, $N2 \geq 2$, and N2 is a positive integer less than or equal to N.

[0016] According to the foregoing embodiment, a second apparatus may deliver active parameters of a plurality of layers at a time, a quantity of times of delivering the active parameters is reduced, and signaling overheads caused by frequent delivery of the active parameters are reduced.

[0017] In some implementations, the first model includes M layers, M is a positive integer greater than N2, and numbers of the N2 layers among the M layers are non-consecutive.

[0018] According to the foregoing embodiment, the first active parameter indicated by the first information may include active parameters of layers with non-consecutive numbers, that is, one or more layers may be skipped for these active parameters, to reduce an amount of information for transmitting the first information and reduce latency.

[0019] In some implementations, N2 is related to a resource for transmitting the first

information and/or an inference latency requirement.

[0020] According to the foregoing embodiment, a value of N2 may be related to the resource for transmitting the first information and/or the inference latency requirement, so that transmission of the first information can meet a resource limitation and/or the inference latency requirement. The foregoing solution can ensure that the first information can be successfully transmitted, or can ensure that transmission of the first information can meet the inference latency requirement of the first model, to further ensure inference performance of the first model.

[0021] In some implementations, the method further includes: predicting the information about the location of the first active parameter based on an input value of a layer preceding a 1st layer in the N2 layers.

[0022] According to the foregoing embodiment, in one aspect, the first apparatus may predict information about locations of active parameters of a plurality of first modules that are respectively included in a plurality of layers. Compared with a solution in which an active parameter of only one module is predicted at a time, the foregoing embodiment can save computing resources. In another aspect, the first apparatus predicts the N2 layers based on the input value of the layer preceding the 1st layer in the N2 layers, instead of based on an input value of the 1st layer in the N2 layers. In this way, the first apparatus may predict information about a location of an active parameter of a first module at the N2 layers when inferring the layer preceding the 1st layer in the N2 layers. In other words, a layer for which the first apparatus predicts the location of the active parameter is not a layer that is performing inference. Therefore, the foregoing embodiment does not affect the layer that is performing inference, that is, does not reduce inference performance of the large model.

[0023] In some implementations, each of the N2 layers further includes one second module, and the second module is one of N2 second modules in the first model. The method further includes: predicting information about a location of a third active parameter based on the input value of the layer preceding the 1st layer in the N2 layers, where the third active parameter is an active parameter of the N2 second modules.

[0024] According to the foregoing embodiment, in one aspect, the first apparatus may predict information about locations of active parameters of a plurality of second modules that are respectively included in a plurality of layers. Compared with a solution in which an active parameter of only one module is predicted at a time, the foregoing embodiment can save computing resources. In another aspect, a layer for which the first apparatus predicts the location of the active parameter is not a layer that is performing inference. Therefore, the foregoing embodiment does not affect the layer that is performing inference, that is, does not reduce inference performance of the large model.

[0025] In some implementations, the second module is an MHA module.

[0026] In some implementations, each of the N2 layers further includes one second module. The method further includes: determining a third input value based on an output value of a first module at a kth layer in the N2 layers and an output value of a second module at the kth layer, where the third input value is used as an input of a first module at a layer following the kth layer and an input of a second module at the layer following the kth layer, and k is a positive integer less than or equal to N2.

[0027] According to the foregoing embodiment, one or more layers among the N2 layers may use a form of parallel inference. That is, when a first module at a layer performs inference, a second module at the layer may also perform inference. Sequential inference is to use an output obtained after inference of a second module at a layer is completed as an input of the second module at the layer. Compared with the layer using sequential inference, the layer using parallel inference can reduce a quantity of inputs from 2 to 1. A quantity of inputs is reduced, and a change rate of adjacent inputs is fixed, so that a value of N2 that can be selected may be increased, and the active parameter indicated by the first information delivered by the second apparatus may be active parameters of more layers. In other words, the second apparatus may deliver the active parameters of more layers at a time, to further reduce a transmission latency from the second apparatus to the first apparatus.

[0028] In some implementations, the method further includes: sending second information when confidence of an output value of the first model is greater than or equal to a first threshold, where the second information indicates stopping transmission of the first information.

[0029] According to the foregoing embodiment, the first apparatus may send the second information when the confidence of the output value of the first model meets a condition, where the second information may indicate stopping transmission of the first information, to reduce a resource waste caused by transmitting the first information. In addition, the foregoing embodiment may be applied to an inference early exiting scenario (early exiting). In the scenario in which the first model exits inference early, the first model exits inference not because inference is completed at all layers in the first model, but because the confidence of the output value of the first model meets the condition. The early exit of inference may enable a part of layers in the first model not to participate in inference, to reduce an inference latency. In the foregoing embodiment, when the first model exits the inference early, the second apparatus can be prevented from transmitting information indicating an active parameter used for inference, to save transmission resources.

[0030] In some implementations, the method further includes: sending third information, where the third information includes indication information of the location of the first active parameter, and the third information is used to request the first active parameter.

[0031] According to the foregoing embodiment, the first apparatus may request delivery of the first active parameter, and the request sent by the first apparatus may carry the indication information of the location of the first active parameter, making a value of the active parameter indicated by the first information more accurate.

[0032] In some implementations, the first module is an FNN module.

[0033] According to a second aspect, a communication method is provided. The method may be performed by a second apparatus (for example, a network device, a core network device, a cloud-side device, or an edge-side device), may be performed by a component (for example, a processor, a chip, or a system on a chip) in the second apparatus, or may be performed by a logical module or software that can implement all or a part of functions of the second apparatus.

[0034] The method includes: generating first information, where the first information indicates a first active parameter, the first active parameter is an active parameter of N first modules in a first model, the first active parameter is used to perform inference based on the N first modules, and N is a positive integer; and sending the first information.

[0035] In some implementations, the N first modules include N1 first modules, the N1 first modules are modules at a first layer of at least one layer in the first model, and N1 is a positive integer less than or equal to N.

[0036] In some implementations, the N first modules further include N2 first modules, the first model includes N2 layers, each of the N2 layers includes one of the N2 first modules, $N2 \geq 2$, and N2 is a positive integer less than or equal to N.

[0037] In some implementations, the first model includes M layers, M is a positive integer greater than N2, and numbers of the N2 layers among the M layers are non-consecutive.

[0038] In some implementations, N2 is related to a resource for transmitting the first information and/or an inference latency requirement.

[0039] In some implementations, the method further includes: receiving second information, where the second information indicates stopping transmission of the first information.

[0040] In some implementations, the method further includes: receiving third information, where the third information includes indication information of location of the first active parameter.

[0041] In some implementations, the sending first information includes: sending the first information in response to the third information.

[0042] In some implementations, the first module is an FNN module.

[0043] According to a third aspect, a communication apparatus is provided, including a processing circuit (or a processor) and an input/output interface (which may also be referred to as an interface circuit), where the input/output interface is configured to input and/or output a signal, and the processing circuit is configured to perform any one of the first aspect and the possible

methods of the first aspect or any one of the second aspect and the possible methods of the second aspect.

[0044] In some implementations, the processing circuit is configured to communicate with another apparatus through the interface circuit, and perform any one of the first aspect and the possible methods of the first aspect or any one of the second aspect and the possible methods of the second aspect.

[0045] According to a fourth aspect, a communication apparatus is provided. The communication apparatus may include a unit, a module, or the like that is configured to perform a function of the communication apparatus.

[0046] In some implementations, the communication apparatus may include a module, a unit, or a means (means) that is in one-to-one correspondence with and configured to perform the method/operation/step/action described in any one of the first aspect and the possible implementations of the first aspect. The module, the unit, or the means may be a hardware circuit, may be software, or may be implemented by a combination of a hardware circuit and software.

[0047] In some implementations, the communication apparatus may include a module, a unit, or a means that is in one-to-one correspondence with and configured to perform the method/operation/step/action described in any one of the second aspect and the possible implementations of the second aspect. The module, the unit, or the means may be a hardware circuit, may be software, or may be implemented by a combination of a hardware circuit and software.

[0048] According to a fifth aspect, a computer-readable storage medium is provided. The computer-readable storage medium stores a computer program or instructions. When the computer program or the instructions are run on a computer, the method according to any one of the first aspect and the possible implementations of the first aspect or the method according to any one of the second aspect and the possible implementations of the second aspect is performed.

[0049] According to a sixth aspect, a computer program product is provided, including a computer program or instructions. When the computer program or the instructions are run on a computer, the method according to any one of the first aspect and the possible implementations of the first aspect or the method according to any one of the second aspect and the possible implementations of the second aspect is performed (or is implemented).

[0050] According to a seventh aspect, a communication apparatus is provided, including a processor, configured to execute a computer program (or computer executable instructions) stored in a memory, and/or a logic circuit, to enable the apparatus to perform any possible method of the first aspect or any possible method of the second aspect.

[0051] In a possible implementation, the apparatus further includes the memory. In a possible

implementation, the processor and the memory are integrated together. In another possible implementation, the memory is located outside the communication apparatus. There may be one or more processors.

[0052] In a possible implementation, the communication apparatus further includes a communication interface. The communication interface is configured for the communication apparatus to communicate with another device, for example, send or receive data and/or a signal. For example, the communication interface may be a transceiver, a circuit, a bus, a module, or another type of communication interface.

[0053] In an implementation, the communication apparatus in the third aspect, the fourth aspect, or the seventh aspect may be a chip or a system on a chip.

[0054] According to an eighth aspect, a chip is provided, including a processor, configured to invoke a computer program or computer instructions in a memory, to enable the processor to perform any implementation of the first aspect or any implementation of the second aspect.

[0055] In an implementation, the processor is coupled to the memory through an interface.

[0056] According to a ninth aspect, a communication system is provided, including a first apparatus and a second apparatus. The first apparatus is configured to perform any one of the first aspect and the possible implementations of the first aspect, and the second apparatus is configured to perform any one of the second aspect and the possible implementations of the second aspect.

[0057] For descriptions of beneficial effect of any one of the second aspect to the ninth aspect, refer to descriptions of beneficial effect of the first aspect.

BRIEF DESCRIPTION OF DRAWINGS

[0058] FIG. 1 is a diagram of a communication system;

[0059] FIG. 2 is a block diagram of another communication system;

[0060] FIG. 3 is a diagram of an inference method;

[0061] FIG. 4 is a schematic flowchart of a communication method according to an embodiment of this application;

[0062] FIG. 5 is a schematic flowchart of another communication method according to an embodiment of this application;

[0063] FIG. 6 is a schematic flowchart of still another communication method according to an embodiment of this application;

[0064] FIG. 7 is a diagram of yet another communication method according to an embodiment of this application;

[0065] FIG. 8A and FIG. 8B are a schematic flowchart of yet another communication method

according to an embodiment of this application;

[0066] FIG. 9 is a diagram of a large model used for parallel inference according to an embodiment of this application;

[0067] FIG. 10 is a block diagram of a communication apparatus according to an embodiment of this application; and

[0068] FIG. 11 is a block diagram of another communication apparatus according to an embodiment of this application.

DESCRIPTION OF EMBODIMENTS

[0069] The following describes technical solutions of this application with reference to accompanying drawings.

[0070] All aspects, embodiments, or features are presented in this application by describing a system that may include a plurality of devices, components, modules, and the like. It should be appreciated and understood that, each system may include another device, component, module, and the like, and/or may not include all devices, components, modules, and the like discussed with reference to the accompanying drawings. In addition, a combination of these solutions may be used.

[0071] In addition, in embodiments of this application, the words such as "example" and "for example" are used to represent giving an example, an illustration, or a description. Any embodiment or design scheme described as an "example" in this application should not be explained as being more preferred or having more advantages than another embodiment or design scheme. Specifically, the word "example" is used to represent a concept in a specific manner.

[0072] A service scenario described in embodiments of this application is intended to describe the technical solutions in embodiments of this application more clearly, but does not constitute a limitation on the technical solutions provided in embodiments of this application. A person of ordinary skill in the art may learn that as a new service scenario emerges, the technical solutions provided in embodiments of this application are also applicable to a similar technical problem.

[0073] Reference to "an embodiment", "some embodiments", or the like described in this specification means that one or more embodiments of this application include a specific feature, structure, or characteristic described with reference to the embodiment. Therefore, statements "in an embodiment", "in some embodiments", "in some other embodiments", "in other embodiments", and the like that appear at different places in this specification do not necessarily mean referring to a same embodiment. Instead, the statements mean "one or more but not all of embodiments", unless otherwise specifically emphasized in another manner. The terms "include", "comprise",

"have", and their variants all mean "including but are not limited to", unless otherwise specifically emphasized in another manner.

[0074] Unless otherwise specified, descriptions such as "first" and "second" in embodiments of this application are merely used for illustration and distinguishing between described objects, do not indicate any order, and do not indicate a particular limitation on a quantity in embodiments of this application, and do not constitute any limitation on embodiments of this application.

[0075] It should be understood that, in embodiments of this application, sequence numbers of processes do not mean execution sequences. The execution sequences of the processes should be determined based on functions and internal logic of the processes, and should not constitute any limitation on implementation processes of embodiments of this application.

[0076] It may be understood that the term "and/or" in this specification describes only an association relationship between associated objects and represents that three relationships may exist. For example, A and/or B may represent the following three cases: Only A exists, both A and B exist, and only B exists. In addition, the character "/" in this specification generally indicates an "or" relationship between the associated objects.

[0077] For ease of understanding this application, the following describes terms that may appear in this application as an example.

[0078] MHA module (block): The MHA module may be a multi-head attention mechanism module in a transformer (transformer) structure of a large model. The large model may be an artificial intelligence (artificial intelligence, AI) model or a machine learning (machine learning, ML) model.

[0079] FNN module (block): The FNN module may be a feedforward neural network module placed behind an MHA module in a large model with a transformer structure. An input of the FNN module may be an output of the MHA module.

[0080] Transformer (transformer) layer: One transformer layer may include an MHA module and at least one FNN module in sequence. An output of an MHA module at a layer may be used as an input of an FNN module at the layer. The large model may include a transformer structure, and one transformer structure includes one or more transformer layers. The following uses an example in which one transformer layer includes one MHA module and one FNN module. However, it should be noted that this application is also applicable to a case in which one transformer layer includes a plurality of MHA modules and a plurality of FNN modules.

[0081] Input (input): The input may be an input of a large model, or an input of a 1st one of a plurality of modules at a first layer in the large model. For example, the input may be a vector or a tensor obtained by performing embedding (embedding) on a request sent by a user to the large model. For another example, the input may be data obtained by the large model from a user side

(for example, a terminal-side device) in an event-triggered or periodic manner. Alternatively, the input may be an input to an intermediate block (for example, the FNN module or the MHA module) in the large model. For example, the input may be a vector or tensor input to a 2nd one of the plurality of modules at the first layer in the large model. For another example, the input may be a vector or a tensor of a module that is input to a second layer or any layer after the second layer in the large model.

[0082] Predictor (predictor) module: The predictor may be an AI model or an ML model. For example, the predictor may be a trained small deep neural network (deep neural network, DNN). The predictor may predict a location of an active parameter of the MHA module or the FNN module based on the input.

[0083] Active parameter (active parameter): The active parameter may include a weight parameter of an active neuron (active neuron) and/or a weight parameter corresponding to an active head (active head).

[0084] Active neuron: An output of the FNN module may be expressed as $\sigma(xW+b)$. σ represents an activation function, where the activation function may be, for example, a rectified linear unit (rectified linear unit, ReLU); x may represent an input (input) to the FNN module; W may represent a weight parameter (for example, a weight matrix); and b may represent a bias parameter (for example, a bias matrix). For the FNN module, the active parameter may be a weight parameter (or referred to as an element in a weight matrix) of an active neuron. The active neuron may be a neuron corresponding to a weight parameter (or referred to as a neuron weight parameter) whose inner product with the input of the FNN module exceeds a preset threshold. For example, if a result obtained by performing calculation on an input x_1 according to x_1W+b is a negative number or a very small positive number, that is, less than the preset threshold, after the result is further processed through the activation function, an obtained output value $\sigma(x_1W+b)$ has little impact on a final output result of the FNN module. Therefore, a related neuron used to calculate x_1 may be an inactive neuron. An element corresponding to the inactive neuron in the matrix W may be set to 0, to obtain a matrix W'. The matrix W' may be a sparse matrix in which most elements are zero, and inference performed based on the matrix W' can save compute resources and reduce an inference latency. For another example, if a result obtained by performing calculation on an input x_2 according to x_2W+b exceeds the preset threshold, it indicates that after the result is further processed through the activation function, an obtained output value $\sigma(x_2W+b)$ has specific impact on a final output result of the FNN module. Therefore, a related neuron used to calculate x_2 is an active neuron. An element corresponding to the active neuron in the matrix W may remain unchanged, so that the active neuron can participate in inference when inference is performed again based on a processed matrix W'.

[0085] Active head (head): An input of the MHA module is multiplied by weight parameters (for example, weight matrices) W_q, W_k, and W_v respectively, to obtain three matrices Q, K, and V. Assuming that a quantity of heads of the MHA module is n, the matrix Q may be divided into n submatrices, which are respectively Q_1 to Q_n, where each submatrix corresponds to one head, and n is a positive integer. For example, the matrix Q is a 512×1 (512 rows and 1 column) matrix. Assuming that n=8, the matrix Q may be divided into eight 64×1 submatrices, which are respectively submatrices Q_1 to Q_8. The submatrices Q_1 to Q_8 each correspond to eight heads. Similarly, the matrix K may be divided into submatrices K_1 to K_n, where each submatrix corresponds to one head; and the matrix V may be divided into submatrices V_1 to V_n, where each submatrix corresponds to one head. In this way, each head corresponds to three submatrices, which are respectively Q_i, K_i, and V_i, and i is an integer ranging from 1 to n. Each head is used to perform inference based on the input and the three submatrices Q_i, K_i, and V_i corresponding to the head, to obtain an output result of the head. An inner product of the input and the output result of the head is calculated. If a result of the inner product exceeds a preset threshold, it indicates that the head is an active head; or otherwise, the head is an inactive head. For the MHA module, the active parameter may be a weight matrix corresponding to at least one active head. For example, an inner product of the input and the output result of the head is calculated. If an obtained result is large, for example, exceeds a preset threshold, the three submatrices Q_i, K_i, and V_i corresponding to the head may be reserved. If an obtained head parameter (or the output result of the head) is small, for example, less than a preset threshold, the three submatrices Q_i, K_i, and V_i corresponding to the head may be set to 0. In this way, for an MHA module including a plurality of heads, submatrices of a part of the plurality of heads may be set to 0, affecting values of the initial weight matrices W_q, W_k, and W_v, and obtaining changed weight matrices W_q', W_k', and W_v'. The changed weight matrices W_q', W_k', and W_v' may be referred to as weight matrices corresponding to the at least one active head.

[0086] It should be noted that "exceeding the preset threshold" may be "greater than or equal to the preset threshold", or may be "greater than the preset threshold"; and "less than the preset threshold" may be "less than the preset threshold", or may be "less than or equal to the preset threshold".

[0087] The technical solutions in embodiments of this application may be applied to various communication systems, including but not limited to: a long term evolution (long term evolution, LTE) system, a 5th generation (5th generation, 5G) mobile communication system such as a new radio (new radio, NR) system, a narrowband internet of things (narrowband internet of things, NB-IoT) system, an enhanced machine type communication (enhanced machine type communication, eMTC) system, an enhanced mobile broadband (enhanced mobile broadband, eMBB) system, an

ultra-reliable and low latency communication (ultra-reliable and low latency communication, URLLC) system, a satellite communication system, an LTE-machine-to-machine (LTE-machine-to-machine, LTE-M) system, or a system evolved after 5G such as a 6th generation (6th generation, 6G) mobile communication system.

[0088] It should be noted that, in embodiments of this application, the term "communication" may also be described as "data transmission", "signal transmission", "information transmission", "transmission", or the like. In embodiments of this application, the transmission may include sending or receiving. For example, the transmission may be uplink transmission, for example, may be that a terminal device sends a signal to a network device; or the transmission may be downlink transmission, for example, may be that a network device sends a signal to a terminal device; or the transmission may be sidelink transmission, for example, may be that a terminal device sends a signal to another terminal device.

[0089] FIG. 1 is a diagram of a communication system 100. As shown in FIG. 1, the communication system 100 includes a radio access network 110 and a core network 120. Optionally, the communication system 100 may further include an internet 130. The radio access network 110 may include at least one network device (for example, 111a and 111b in FIG. 1), and may further include at least one terminal device (for example, 112a to 112j in FIG. 1). The terminal device is connected to the network device in a wireless manner. The network device is connected to the core network 120 in a wireless or wired manner. The core network 120 may include one or more core network devices. The core network device and the network device may be different independent physical devices, or functions of the core network device and logical functions of the network device may be integrated into a same physical device, or a part of functions of the core network device and a part of functions of the network device may be integrated into one physical device. Terminal devices may be connected to each other in a wired or wireless manner and network devices may be connected to each other in a wired or wireless manner. Wireless communication may be performed between the terminal devices, between the network devices, and between the terminal device and the network device through an air interface resource. For example, the air interface resource may include at least one of a time-domain resource, a frequency-domain resource, a code resource, and a spatial resource. It should be noted that FIG. 1 is merely a diagram. The communication system 100 may further include other network devices, for example, a wireless relay device and a wireless backhaul device that are not shown in FIG. 1.

[0090] The network device may be any device having a wireless transceiver function. For example, the network device may be a base station configured to connect the terminal device to the radio access network (radio access network, RAN). The network device may also be referred to as an access network device or an access network node sometimes. It may be understood that,

devices having functions of the network device may have different names in systems using different radio access technologies. For ease of description, in embodiments of this application, an apparatus that provides a wireless communication and access function for the terminal device is collectively referred to as a base station. In embodiments of this application, the network device includes but is not limited to a macro base station (111a in FIG. 1), a micro base station or an indoor station (111b in FIG. 1), a pico base station, a small cell, a balloon station, a relay station, an access point, and the like in various forms. The network device may include an evolved NodeB (evolved NodeB, eNB or eNodeB) in LTE, an access point (Access Point, AP) in a wireless fidelity (wireless fidelity, Wi-Fi) system, a wireless relay node, a wireless backhaul node, a transmission point (transmission point, TP), or a transmission reception point (transmission reception point, TRP), or the like, or may include a next generation base station NodeB (next generation NodeB, gNB) or a transmission point (TRP or TP) in a 5G system, one or a group of antenna panels (including a plurality of antenna panels) of a base station in the 5G system, a network node that forms a gNB or a transmission point, for example, a baseband unit (baseband unit, BBU) or a distributed unit (distributed unit, DU), or may include a network device, a server, a vehicle-mounted device, or the like in a network evolved after 5G such as 6G. Alternatively, the network device may be a module or a unit that completes a part of functions of the base station, for example, may be a central unit (central unit, CU), or may be a DU.

[0091] In embodiments of this application, an apparatus configured to implement a function of the network device may be a network device, or may be an apparatus that can support the network device in implementing the function, for example, a system on a chip. The apparatus may be installed in the network device. The system on a chip may include a chip, or may include a chip and another discrete component.

[0092] In another possible scenario, a plurality of network devices collaborate to assist the terminal in implementing radio access, and different network devices separately implement a part of functions of the base station. For example, the network device may be a CU, a DU, a CU-control plane (control-plane, CP), a CU-user plane (user-plane, UP), or a radio unit (radio unit, RU). The CU and the DU may be separately disposed, or may be included in a same network element, for example, a BBU. The RU may be included in a radio frequency device or a radio frequency unit, for example, included in a remote radio unit (remote radio unit, RRU), an active antenna unit (active antenna unit, AAU), or a remote radio head (remote radio head, RRH).

[0093] In different systems, the CU (or the CU-CP and the CU-UP), the DU, or the RU may also have different names, but a person skilled in the art may understand meanings thereof. For example, in an open radio access network (open radio access network, O-RAN) system, the CU may also be referred to as an O-CU (open CU), the DU may also be referred to as an O-DU, the

CU-CP may also be referred to as an O-CU-CP, the CU-UP may also be referred to as an O-CU-UP, and the RU may also be referred to as an O-RU. For ease of description, the CU, the CU-CP, the CU-UP, the DU, and the RU are used as examples for description in this application. Any one of the CU (or the CU-CP or the CU-UP), the DU, and the RU in this application may be implemented by a software module, a hardware module, or a combination of a software module and a hardware module. A specific technology and a specific device form that are used by the network device are not limited in embodiments of this application.

[0094] The terminal device may be a device that provides voice and/or data connectivity for a user, or the terminal device may be a device having a wireless connection function. The terminal device may be deployed on land, including an indoor or outdoor device, a handheld device, or a vehicle-mounted device; may be deployed on a water surface (for example, on a ship); or may be deployed in the air (for example, on an airplane, a balloon, or a satellite). The terminal device may also be referred to as user equipment (user equipment, UE), an access terminal, a terminal, a subscriber unit (subscriber unit), a subscriber station, a mobile station (mobile station, MS), a mobile terminal (mobile terminal, MT), a remote station, a remote terminal, a mobile device, a user terminal, a wireless network device, a user agent, or a user apparatus. In embodiments of this application, the terminal device includes but is not limited to: a cellular phone (cellular phone), a mobile phone (mobile phone), a wireless data card, a wireless modem (modem), a tablet computer (pad), a laptop computer (laptop computer), a notebook computer, a palmtop computer, a mobile internet device (mobile internet device, MID), a computer having a wireless transceiver function, a cordless phone, a session initiation protocol (session initiation protocol, SIP) phone, a smartphone (smartphone), a wireless local loop (wireless local loop, WLL) station, a personal digital assistant (personal digital assistant, PDA), a handheld device (handset) having a wireless communication function, a compute device or another device connected to a wireless modem, a vehicle-mounted device (for example, a car, a bicycle, an electric vehicle, an airplane, a ship, a train, or a high-speed train), a wearable device (for example, a smart watch, a smart band, a pedometer, or smart glasses), a satellite terminal, a terminal device in the internet of things or the internet of vehicles, a terminal in any form in a future network, relay user equipment, a terminal in a future evolved public land mobile network (public land mobile network, PLMN), or the like. The terminal device may alternatively be a virtual reality (virtual reality, VR) device, an augmented reality (augmented reality, AR) device, a smart point of sale (point of sale, POS) machine, a customer-premises equipment (customer-premises equipment, CPE), a light terminal device (light UE), reduced capability user equipment (reduced capability UE, REDCAP UE), a machine type communication (machine type communication, MTC) terminal, a terminal device in industrial control (industrial control), a terminal device in self-driving (self-driving), a terminal device in

telemedicine (remote medical), a terminal device in smart grid (smart grid), a wireless terminal device in transportation safety (transportation safety), a terminal device in smart city (smart city), a terminal device in smart home (smart home), a tactile terminal device, a smart home device (for example, a refrigerator, a television, an air conditioner, or an electric meter), a smart robot, a robot arm, a workshop device, a wireless terminal in self-driving, or a flight device (for example, a smart robot, a hot air balloon, an uncrewed aerial vehicle, or an airplane), or the like. The terminal device may alternatively be a vehicle apparatus, for example, an entire vehicle apparatus, a vehicle-mounted module, a vehicle-mounted chip, an on-board unit (on-board unit, OBU), or a telematics box (telematics box, T-BOX) of the internet of vehicles; or the terminal device may be another device having a terminal function. For example, the terminal device may be a device that functions as a terminal in device-to-device (device-to-device, D2D) communication. This is not limited in embodiments of this application.

[0095] In embodiments of this application, an apparatus configured to implement a function of the terminal device may be a terminal device, or may be an apparatus that can support the terminal device in implementing the function, for example, a chip or a system on a chip. The apparatus may be installed in the terminal device. The system on a chip may include a chip, or may include a chip and another discrete component. In the technical solutions in embodiments of this application, the technical solutions provided in embodiments of this application are described by using an example in which the apparatus configured to implement the function of the terminal device is the terminal device, which may also be referred to as a terminal (terminal), and the terminal device may be UE.

[0096] Roles of the base station and the terminal may be relative. For example, a helicopter or an uncrewed aerial vehicle 112i in FIG. 1 may be configured as a mobile base station. For the terminal 112j that accesses the radio access network 110 through 112i, the terminal 112i is a base station. However, for the base station 111a, 112i is a terminal. In other words, communication between 111a and 112i is performed based on a radio air interface protocol. Certainly, communication between 111a and 112i may alternatively be performed based on an interface protocol between base stations. In this case, for 111a, 112i is also a base station. Therefore, the base station and the terminal may be collectively referred to as a communication apparatus, 111a and 111b in FIG. 1 may be referred to as communication apparatuses having a base station function, and 112a to 112j in FIG. 1 may be referred to as communication apparatuses having a terminal function.

[0097] The network device may communicate with the terminal device through a radio link. A transmission link from the network device to the terminal device may be referred to as a downlink (downlink, DL) or a downlink channel, and is configured to transmit a downlink signal. A

transmission link from the terminal device to the network device may be referred to as an uplink (uplink, UL) or an uplink channel, and is configured to transmit an uplink signal.

[0098] FIG. 2 is a block diagram of another communication system 200. The communication system 200 may also be referred to as a cloud system or another name.

[0099] The communication system 200 includes a central cloud 210, an edge cloud 220, and a terminal side 230. The edge cloud 220 may be different from the central cloud 210 (or referred to as a conventional cloud platform). For example, the central cloud 210 (or referred to as the conventional cloud platform) may include a hyperscale and concentrated data center, while the edge cloud 220 may include a plurality of edge nodes distributed in different regions, where a resource scale of a single edge node is small. The edge cloud 220 covers a wider network range, and is closer to a terminal-side device in the terminal side 230. The terminal-side device in the terminal side 230 and the edge node in the edge cloud 220 may communicate with each other through a routing system or in another manner.

[00100] The edge node may externally provide various resources, for example, a compute resource having a computing capability, such as a central processing unit (central processing unit, CPU), a graphics processing unit (graphics processing unit, GPU), or a tensor processing unit (tensor processing unit, TPU), for another example, a storage resource having a storage function, such as a memory or a hard disk drive, and a network resource such as a bandwidth. Types and quantities of resources included in two of the plurality of edge nodes may be the same, or may be different. The plurality of edge nodes may be deployed by a same internet service provider, or may be provided by different internet service providers. A service that provides cloud computing can be deployed on the edge node in the edge cloud 220, that is, an application that can provide a service is deployed on the edge node. For example, a large model function can be deployed on the edge node.

[00101] The central cloud 210 may also be referred to as a cloud-side device. The edge cloud 220 or the edge node in the edge cloud 220 may also be referred to as an edge-side device. For example, the terminal-side device may be a terminal device, for example, the terminal device shown in FIG. 1.

[00102] In some possible implementations, the central cloud 210 may be deployed in the core network 120. In other words, the core network 120 may have a function of the central cloud 210. In some possible implementations, the edge cloud 220 may be deployed in a network device. Alternatively, the network device may include an edge node.

[00103] The rapid development of large models has led to the emergence of applications in various industries. A cloud-side large model may be a large model deployed on a cloud-side device or an edge-side device. A terminal-side large model may be a large model deployed on a terminal-

side device. The terminal-side large model can meet user requirements for protecting terminal-side privacy data, thereby attracting widespread attention. However, due to limited storage space, the terminal-side device can only store a large model with a small parameter quantity, which performs much worse than a large model with a large parameter quantity, hindering the application of the large model on the terminal side.

[00104] FIG. 3 is a diagram of an inference method. A large model shown in FIG. 3 may be deployed on a terminal side, and the large model may use a transformer structure.

[00105] The right part of FIG. 3 is a legend of the left part. The left part of FIG. 3 shows an inference process of the first three layers of the large model. Each layer is represented by a dashed rounded rectangle, and each layer includes one MHA module and one FNN module. For example, the MHA module may include four heads, and the four heads may be represented by four rectangles in FIG. 3. An input of the MHA module is represented by a dashed (dashed line) arrow in FIG. 3. The FNN module may include five neurons, and the five neurons are represented by five squares. An input of the FNN module is represented by a dotted (dotted line) arrow. A predictor module may predict a location of an active parameter of the MHA module or the FNN module based on the input, and a prediction process is represented by a curved solid arrow. P1 to P4 in FIG. 3 respectively represent the first to fourth predictions of the predictor module. For example, the first prediction is a prediction performed on the active parameter of the MHA module at the second layer based on an output of the FNN module at the first layer.

[00106] Specifically, a first input may be an input to the MHA module at the first layer. A terminal-side device may perform inference on an input of a user based on full parameters and the MHA module at the first layer, to obtain an output result. In other words, although not all heads in the MHA module at the first layer are active heads, when the MHA module at the first layer performs inference, submatrices of all heads need to be loaded into a memory of the terminal-side device. The output result of the MHA module at the first layer may be used as an input of the FNN module at the first layer. The terminal-side device may perform inference on the input based on full parameters and the FNN module at the first layer, to obtain an output result. In other words, although not all neurons in the FNN module at the first layer are active neurons, when the FNN module at the first layer performs inference, the terminal-side device needs to load all weight parameters to the memory of the terminal-side device. When the FNN module at the first layer performs inference, the predictor module may predict a location of the active parameter of the MHA module at the second layer based on the input of the FNN module at the first layer, which is represented by P1 and a curved solid arrow related to P1 in FIG. 3.

[00107] The output result of the FNN module at the first layer may be used as an input of the MHA module at the second layer. The terminal-side device obtains an active parameter from a

flash memory of the terminal-side device based on the location of the active parameter of the MHA module at the second layer that is predicted by the predictor module, and performs inference on the input based on the active parameter and the MHA module at the second layer, to obtain an output result. In other words, in the MHA module at the second layer, only the active parameter of the active head is loaded into the memory. For example, heads at a first location and a third location from the left in FIG. 3 are active heads of the MHA module at the second layer. Correspondingly, weight matrices corresponding to the two active heads are the active parameters. In addition, because the terminal-side device cannot predict which parameters in the full parameters are the active parameters, the flash memory of the terminal-side device needs to store the full parameters of the MHA module, to determine the active parameters in the full parameters based on the locations of the active parameters. When the MHA module at the second layer performs inference, the predictor module may predict the active parameter of the FNN module at the second layer based on the input of the MHA module at the second layer, which is represented by P2 and a curved solid arrow related to P2 in FIG. 3. The output result of the MHA module at the second layer may be used as an input of the FNN module at the second layer. The terminal-side device may obtain an active parameter from the flash memory of the terminal-side device based on the location of the active parameter of the FNN module at the second layer that is predicted by the predictor module, and perform inference on the input based on the active parameter and the FNN module at the second layer, to obtain an output result. In other words, in the FNN module at the second layer, only the weight parameter of the active neuron is loaded into the memory. For example, neurons at a third location and a fifth location from the left in FIG. 3 are active neurons. Correspondingly, elements in weight matrices corresponding to the two neurons are the active parameters. In addition, because the terminal-side device cannot predict which parameters in the full parameters are the active parameters, the flash memory of the terminal-side device needs to store the full parameters of the FNN module, to determine the active parameters in the full parameters based on the locations of the active parameters. When the FNN module at the second layer performs inference, the predictor module may predict the active parameter of the MHA module at the third layer based on the input of the FNN module at the second layer, which is represented by P3 and a curved solid arrow related to P3 in FIG. 3. For descriptions of the MHA module and the FNN module at the third layer, refer to the foregoing descriptions of the second layer. Details are not described again.

[00108] The terminal-side device usually has small flash memory space. According to the solution shown in FIG. 3, in a process of performing inference, the terminal-side device needs to read the required active parameter from the flash memory. Therefore, the terminal-side device needs to store all the weight matrices of the MHA module and all the weight parameters of the

FNN module in the flash memory. Limited by flash memory space, the terminal-side device can only store a large model with a small parameter quantity. For example, a large model with 7 billion (billion, B) parameters is stored in a single-precision floating-point format, and requires flash memory space of 7 B×32 bits (bit)/8 bits per byte (byte)=28 gigabytes (GB). The terminal-side device cannot store a large model with a large parameter quantity. For example, the terminal-side device cannot store a large model with a quantity of more than 10 billion parameters. A person skilled in the art may understand that a large model with a small parameter quantity performs much worse than a large model with a large parameter quantity, and consequently the terminal-side large model cannot achieve performance comparable to that of a cloud-side large model, affecting user experience.

[00109] The terminal-side device usually has small memory space. According to the solution shown in FIG. 3, the first transformer layer uses full parameters for inference. For example, a generative pre-trained transformer (generative pre-trained transformer, GPT) 3 includes 96 transformer layers, and space occupied by full parameters of each transformer layer is 175 B/96×4=7.3 GB. In other words, when GPT3 performs inference of the first transformer layer, all the parameters of 7.3 GB need to be loaded into the memory. However, most terminal-side devices with limited memory space cannot cope with such high memory usage. For example, a memory of a mobile phone may be only 8 GB. Limited memory space degrades performance of the terminal-side large model, and consequently the terminal-side large model cannot achieve performance comparable to that of a cloud-side large model, affecting user experience.

[00110] Therefore, in an inference process of the terminal-side large model, how to cope with a contradiction between an excessively large parameter quantity of the large model and excessively small storage space, to improve inference performance of the terminal-side large model is an urgent problem to be resolved.

[00111] FIG. 4 is a schematic flowchart of a communication method 400 according to an embodiment of this application. The method 400 can improve inference performance of a terminal-side large model. The following describes the method 400 with reference to FIG. 4.

[00112] S430: A first apparatus receives first information from a second apparatus. Correspondingly, the second apparatus sends the first information to the first apparatus.

[00113] For example, the first apparatus may be a terminal device or a terminal-side device. In S430, the first information may be received by the first apparatus, may be received by a component (for example, a processor, a chip, or a system on a chip) in the first apparatus, or may be received by a logical module or software that can implement all or a part of functions of the first apparatus.

[00114] For example, the second apparatus may be a network device, a core network device, a cloud-side device, or an edge-side device. In S430, the first information may be sent by the second

apparatus, may be sent by a component (for example, a processor, a chip, or a system on a chip) in the second apparatus, or may be sent by a logical module or software that can implement all or a part of functions of the second apparatus.

[00115] Similar to the foregoing descriptions, other operations related to the first apparatus and the second apparatus in this application may also be performed by the apparatus, a component in the apparatus, or a logical module or software that can implement all or a part of functions of the apparatus. Details are not described again.

[00116] Optionally, the first information indicates a first active parameter. The first active parameter may be an active parameter of N first modules in a first model, and N is a positive integer.

[00117] The first model may be any large model, AI model, or ML model. For example, the first model may be GPT1, GPT2, GPT3, or the like. The first module may be any module in the first model. For example, the first module may be an FNN module. Therefore, the first active parameter may be an active parameter of the FNN module. However, this is not limited in this application. For example, the first module may be an MHA module or another module. In this way, the first active parameter may be an active parameter of the MHA module or another module. The first model may include the N first modules. In other words, the first model may include one or more first modules. Therefore, the first active parameter may be an active parameter of one or more modules.

[00118] A name of the first model is not limited in this application. The first model may also be referred to as an AI model, an ML model, or another name. A name of the first active parameter is not limited in this application. The first active parameter may also be referred to as a parameter, an active neuron, an active weight, an active element, or another name.

[00119] It should be noted that a location of the first module is not limited in this application, or a quantity of layers at which the first module is located is not limited. For example, the first module may be located at a first layer. For another example, the first module may be located at a third layer. For still another example, three first modules may be respectively located at a second layer, a third layer, and a fourth layer. Therefore, this application does not limit which layer or layers the first active parameter belongs to. For example, the first active parameter may be an active parameter of a module at the first layer (referred to as an active parameter at the first layer for short, and other descriptions are similar). For another example, the first active parameter may be an active parameter at the third layer. For another example, the first active parameter may include an active parameter at the second layer, an active parameter at the third layer, and an active parameter at the fourth layer.

[00120] A manner in which the first information indicates the first active parameter is not

limited in this application. In an example, the first information may include the first active parameter. In this way, the first apparatus may obtain the first active parameter based on content of the first information. In another example, the first information may include an identifier of the first active parameter. The first apparatus may pre-obtain a correspondence between a plurality of active parameters and a plurality of identifiers, for example, a mapping table between the active parameters and the identifiers. The first apparatus may determine, based on the identifier of the first active parameter and the pre-obtained correspondence, that the active parameter corresponding to the identifier is the first active parameter, to obtain the first active parameter.

[00121] A name of the first information is not limited in this application. The first information may also be referred to as active parameter information or another name.

[00122] In some possible implementations, the method 400 further includes: (S420): The second apparatus generates the first information. In some possible implementations, S420 is performed before S430. In some possible implementations, S420 may be replaced with the following: The second apparatus obtains the first information. For example, the second apparatus is a radio frequency chip in a base station, and the radio frequency chip may obtain the first information from a baseband chip in the base station.

[00123] A manner in which the second apparatus generates the first information is not limited in this application. In an example, the second apparatus may prestore all or a part of active parameters. The second apparatus may determine the first active parameter based on the prestored active parameters, and generate the first information indicating the first active parameter. In another example, the second apparatus may obtain the first active parameter from another apparatus, and generate the first information indicating the first active parameter. The another apparatus may be an apparatus different from the first apparatus and the second apparatus.

[00124] A trigger condition for generating the first information by the second apparatus is not limited in this application. In an example, the second apparatus may generate the first information based on a request of the first apparatus. For example, the second apparatus may receive request information from the first apparatus, where the request information indicates the second apparatus to deliver the first active parameter; and the second apparatus may generate the first information based on the request information. In another example, the second apparatus may generate the first information based on a request of another apparatus. For example, the second apparatus may receive request information from the another apparatus, where the request information indicates the second apparatus to send the first active parameter to the first apparatus; and the second apparatus may generate the first information based on the request information. In still another example, the second apparatus may actively generate the first information. In other words, the second apparatus may autonomously determine to generate the first information instead of based

on a request of an apparatus.

[00125] In some possible implementations, the method 400 further includes: (S440): The first apparatus performs inference based on the N first modules according to the first active parameter.

[00126] Optionally, the first active parameter includes an active parameter of one FNN module. The first apparatus may set an element at a location of the active parameter in a weight matrix of the FNN module to the active parameter indicated by the first information, and set an element at another location in the weight matrix to 0, to obtain a processed weight matrix. The first apparatus may perform inference based on an input of the FNN module and the FNN module for which the processed weight matrix is configured.

[00127] It should be noted again that one FNN module is used as an example above. However, that the first active parameter can include the active parameter of only one FNN module is not limited in this application. The first active parameter may further include active parameters of more FNN modules.

[00128] According to the foregoing embodiment, the first apparatus for inference does not need to store full parameters of the first module, but may obtain active parameters in the full parameters of the first module through the first information. Compared with a solution in which the full parameters are stored locally in the first apparatus and the active parameters are read from the full parameters, the foregoing embodiment can save storage space for the first apparatus, so that the first apparatus with limited storage space can run a large model with a large parameter quantity, thereby improving inference performance of the large model. For example, the first module may be an FNN module in a large model with a transformer structure. A ratio of full parameters of the FNN module to full parameters of an MHA module is about 2:1. If the full parameters of the FNN module are not stored in the terminal-side device, two thirds of flash memory space of the terminal-side device is saved. The saved flash memory space may store a parameter of another module, for example, the parameter of the MHA module. In this way, a quantity of parameters of the large model can be increased by three times, thereby effectively improving performance of the large model.

[00129] Optionally, the first active parameter includes an active parameter of one or more MHA modules. Optionally, the first active parameter includes an active parameter of one or more MHA modules and an active parameter of one or more FNN modules.

[00130] In some possible implementations, the method 400 further includes: (S410): The first apparatus sends third information to the second apparatus. Correspondingly, the second apparatus receives the third information from the first apparatus.

[00131] Optionally, the third information includes indication information of a location of the first active parameter. The indication information of the location of the first active parameter may

indicate the location of the first active parameter. In an example, the indication information of the location of the first active parameter may include a specific location of the first active parameter. For example, the location of the first active parameter may be a location of an element in a weight matrix of the FNN module. For example, for a 2×4 weight matrix, a first row and a second column and a second row and a second column may be the location of the first active parameter. In another example, the indication information of the location of the first active parameter may include an identifier of the location of the first active parameter. The first apparatus may pre-obtain a correspondence between locations of a plurality of active parameters and a plurality of identifiers. For example, the first apparatus may pre-obtain a mapping table, where the mapping table includes a mapping between the locations of the plurality of active parameters and the plurality of identifiers. The first apparatus may determine, based on the identifier of the location of the first active parameter, the location of the first active parameter corresponding to the identifier. For example, an identifier #1 corresponds to a first row and a second column and a second row and a second column, so that the first apparatus can determine, based on the identifier #1 indicated by the indication information of the location of the first active parameter, that the location of the first active parameter is the first row and the second column and the second row and the second column.

[00132] For another example, the location of the first active parameter may be a location of an active head of the MHA module. For example, for an MHA module having eight heads, a third head and a fourth head in the eight heads may be active heads. In this way, the indication information of the location of the first active parameter may indicate the third head and the fourth head. A manner of indicating the third head and the fourth head may be direct indication, or may be indirect indication. Details are not described again.

[00133] Optionally, the third information is used to request the first active parameter. In other words, the third information may be used to request the second apparatus to deliver the first active parameter. Alternatively, the third information may be used to request the second apparatus to send the first active parameter to the first apparatus.

[00134] According to the foregoing embodiment, the first apparatus may request delivery of the first active parameter, and the request sent by the first apparatus may carry the indication information of the location of the first active parameter, making a value of the active parameter indicated by the first information more accurate.

[00135] In some possible implementations, S430 includes: The second apparatus sends the first information to the first apparatus in response to the third information.

[00136] In other words, the second apparatus may send the first active parameter to the first apparatus based on the request of the first apparatus.

[00137] It should be noted that this application does not restrict the first apparatus to reporting

the indication information of the location of the first active parameter to the second apparatus. In some other possible implementations, the second apparatus may obtain the indication information of the location of the first active parameter from another apparatus. The indication information of the location of the first active parameter may be direct indication information, that is, the indication information includes the location information; or the indication information of the location of the first active parameter may be indirect indication information, and the second apparatus may determine the location information of the first active parameter based on the indication information.

[00138] In some implementations, the method 400 further includes: (S450): The first apparatus sends second information to the second apparatus when confidence of an output value of the first model is greater than or equal to a first threshold. Correspondingly, the second apparatus receives the second information from the first apparatus.

[00139] For example, the confidence of the output value of the first model may be calculated according to a cosine similarity of output values of two adjacent layers in the first model. However, a manner of calculating the confidence of the output value of the first model is not limited in this application. The confidence of the output value of the first model may alternatively be calculated in another manner. For example, a dedicated linear classifier may be trained based on a small AI model or ML model, and the dedicated linear classifier may calculate the confidence of the output value of the first model by using the output value of the first model. The first threshold may be preset. For example, the first threshold may take a large value. For example, the first threshold may be 95%. A name of the first threshold is not limited in this application. The first threshold may also be referred to as a confidence threshold, a preset threshold, a stop threshold, or another name.

[00140] Optionally, the second information indicates stopping transmission of the first information. In an example, when the first apparatus has requested the first active parameter, but the first apparatus has not received the first information, the first apparatus may perform S450, that is, the first apparatus may send the second information to the second apparatus when a confidence condition is met, to indicate that the first apparatus no longer transmits information indicating the first active parameter. In another example, when the first apparatus receives a part of the first information, for example, the first apparatus receives information indicating active parameter of a part of N2 first modules, the first apparatus may perform S450, that is, the first apparatus may send the second information to the second apparatus when a confidence condition is met, to indicate that the first apparatus stops transmission of the first information, that is, no longer transmits the remaining part of the first information.

[00141] A name of the second information is not limited in this application. The second information may also be referred to as stop information or another name.

[00142] In some possible implementations, the N first modules include N1 first modules, the N1 first modules are modules at a first layer of at least one layer in the first model, and N1 is a positive integer less than or equal to N.

[00143] The first model may include the at least one layer. For example, the first model may include at least one transformer layer. The N1 first modules may be modules at the first layer of the at least one layer. For example, N1=1, and the first module may be an FNN module at a first transformer layer in the first model. For another example, N1=1, and the first module may be an MHA module at a first transformer layer in the first model.

[00144] If the N first modules include N1 first modules, the first active parameter may include an active parameter of the N1 first modules in the first model. If the N1 first modules may be modules at a first layer, the first active parameter may include an active parameter at the first layer in the first model. With reference to S430, the second apparatus may send the active parameter at the first layer in the first model to the first apparatus. With reference to S440, the first apparatus may perform inference based on the model at the first layer according to the active parameter instead of the full parameters.

[00145] According to the foregoing embodiment, the first apparatus may perform inference based on the module at the first layer in the first model according to the active parameter indicated by the first information. Compared with a solution in which inference is performed based on the full parameters and the module at the first layer, this embodiment of this application does not require the full parameters of the module at the first layer to be loaded into a memory, so that memory space of the first apparatus is saved, thereby further improving inference performance of the first model.

[00146] FIG. 5 is a schematic flowchart of another communication method 500 according to an embodiment of this application. The method 500 may be combined with the method 400. In the method 500, full parameters of a 1st one of a plurality of modules at a first layer in a first model may be stored locally in a first apparatus, and full parameters of a 2nd one of the plurality of modules at the first layer in the first model may be stored in a second apparatus. In other words, the first model needs to request an active parameter of the 2nd one of the plurality of modules at the first layer from the second apparatus. The second apparatus may determine an active parameter (that is, a first active parameter) based on the full parameters of the 2nd one of the plurality of modules at the first layer. The following describes the method 500 with reference to FIG. 5.

[00147] In FIG. 5, it is assumed that a first module includes a 2nd one of a plurality of modules at one layer in the first model. It is assumed that a second module includes a 1st one of the plurality of modules at one layer in the first module. However, it should be noted that the description in FIG. 5 is merely used as an example, and does not constitute a limitation on this application. For

example, the first module may include the 1st one and the 2nd one of the plurality of modules at one layer in the first model. In other words, in some other embodiments of this application, the first module may include all modules at one layer in the first model.

[00148] S510: The first apparatus predicts information about a location of a second active parameter based on a first input value. It should be noted that, in FIG. 5, an operation of S510 is shown by using a dashed line, indicating that S510 is an optional operation when S510 is combined with the method 400.

[00149] Optionally, N1 first modules are a 2nd one of the plurality of modules at the first layer in the first model, and N1=1. For example, the first model is a large model with a transformer structure. A 1st one of a plurality of modules at a first transformer layer in the first model is an MHA module, and a 2nd one of the plurality of modules is an FNN module. In other words, the first module may be the FNN module at the first transformer layer in the large model.

[00150] Optionally, the second active parameter in S510 is an active parameter of the 1st one of the plurality of modules at the first layer in the first model. For example, the first module may be the 2nd one of the plurality of modules at the first layer in the first model, and the second active parameter may be an active parameter of a previous module of the first module at the first layer in the first model. For another example, the 1st one of the plurality of modules at the first layer in the first model may be an MHA module. In other words, the second active parameter may be an active parameter of the MHA module at the first layer in the first model.

[00151] A name of the second active parameter is not limited in this application. The second active parameter may also be referred to as a parameter, an active head, an active weight, an active matrix, or another name.

[00152] Optionally, the first input value is an input value of the 1st one of the plurality of modules at the first layer in the first model. For example, the first input value may be an input value of the MHA module at the first transformer layer in the first model. For another example, the first input value may be a vector or a tensor obtained by performing embedding (embedding) on a request sent by a user to a large model.

[00153] For example, the first apparatus may perform prediction by using a predictor module. Specifically, the predictor module may predict the information about the location of the second active parameter based on the first input value. Optionally, the predictor module is deployed in the first apparatus. In other words, the first apparatus includes the predictor module. For example, the predictor module may be a second model. For example, the second model may be an AI model, an ML model, or the like. For another example, the second model may be a DNN model.

[00154] In some possible implementations, the method 500 further includes: (S520): The first apparatus determines the second active parameter based on the information about the location of

the second active parameter. For example, the first apparatus may store full parameters of the first (for example, the MHA module) of the plurality of modules at the first layer in the first model. In this way, the first apparatus may determine a value of the second active parameter based on the information about the location of the second active parameter and the stored full parameters. In another possible implementation, the first apparatus sends request information to the second apparatus; and the second apparatus receives the request information. The request information may be used to request the second active parameter. In this case, the second active parameter may be obtained from the second apparatus. The second apparatus may prestore the full parameters of the first (for example, the MHA module) of the plurality of modules at the first layer in the first model, to determine the requested active parameter from the stored full parameters based on the request information of the first apparatus.

[00155] In some possible implementations, the method 500 further includes: (S530): The first apparatus performs inference based on the 1st one of the plurality of modules at the first layer in the first model according to the second active parameter. For example, the second active parameter is the active parameter of the MHA module at the first layer in the first model. The first apparatus may update a weight matrix of the MHA module by using the second active parameter. Specifically, a weight submatrix of an inactive head in the MHA module is set to 0, and a weight submatrix of an active head is not affected. The first apparatus may perform inference based on the first input value according to the second active parameter (that is, the updated non-zero weight submatrix of the MHA module or the updated weight matrix of the MHA module).

[00156] For example, the method 500 may be applied to a scenario of chatting with a digital intelligent person. For example, a user may send a text message, a picture message, or a voice message to a virtual digital intelligent person on a mobile phone, and receive a text message, a picture message, or a voice message replied by the digital intelligent person. In the scenario of chatting with the digital intelligent person, the virtual digital intelligent person may be driven by the first model. For example, a message sent by the user to the virtual digital intelligent person may be input into the first model. After performing inference based on the foregoing input, the first model may understand and output a message used to reply to the user. In the scenario of chatting with the digital intelligent person, data input into the first model may involve information such as a chat record of the user. Therefore, a large model may be deployed in a terminal-side device such as a mobile phone, and data of the terminal-side device is not uploaded, which protects user privacy. In addition, storage space (for example, a memory or a flash memory) for the terminal-side device such as the mobile phone is limited, and it is difficult to deploy a large model with a large parameter quantity. In addition, when the user chats with the virtual digital intelligent person, a latency requirement for the virtual digital intelligent person to reply to a message is not

high, that is, the user does not require the virtual digital intelligent person to reply to the message immediately.

[00157] According to the foregoing embodiment, the first apparatus may predict the active parameter of the 1st one of the plurality of modules at the first layer in the first model, to perform inference based on the 1st one of the plurality of modules according to the active parameter of the 1st one of the plurality of modules. Compared with a solution in which inference is performed based on the full parameters and the 1st one of the plurality of modules at the first layer, this embodiment of this application does not require the full parameters of the 1st one of the plurality of modules at the first layer to be loaded into a memory, so that memory space of the first apparatus is saved, thereby further improving inference performance of the first model.

[00158] S540: The first apparatus predicts information about a location of first active parameter based on a second input value. It should be noted that, in FIG. 5, an operation of S540 is shown by using a dashed line, indicating that S540 is an optional operation when S540 is combined with the method 400.

[00159] Optionally, the second input value may be an input value of the 2nd one of the plurality of modules at the first layer in the first model or an output value of the 1st one of the plurality of modules at the first layer in the first model. For example, the second input value may be an output value of the MHA module at the first transformer layer in the first model. For another example, the second input value may be an input value of the FNN module at the first transformer layer in the first model. For example, the 2nd one of the plurality of modules at the first layer in the first model may be the FNN module. For example, the 2nd one of the plurality of modules at the first layer in the first model may be the foregoing first module.

[00160] For example, the first apparatus may perform prediction by using a predictor module. Specifically, the predictor module may predict the information about the location of the first active parameter based on the second input value. Optionally, the predictor module is deployed in the first apparatus. In other words, the first apparatus includes the predictor module. For example, the predictor module may be a second model. For example, the second model may be an AI model, an ML model, or the like. For another example, the second model may be a DNN model.

[00161] In some possible implementations, S410 is performed after S540. For example, indication information of the location of the first active parameter included in third information may be predicted by the first apparatus based on the second input value.

[00162] According to the foregoing embodiment, the first apparatus may predict an active parameter of the 2nd one of the plurality of modules at the first layer in the first module, to perform inference based on the 2nd one of the plurality of modules according to the active parameter of the 2nd one of the plurality of modules. Compared with a solution in which inference is performed

based on the full parameters and the 2nd one of the plurality of modules at the first layer, this embodiment of this application does not require the full parameters of the 2nd one of the plurality of modules at the first layer to be loaded into a memory, so that memory space of the first apparatus is saved, thereby further improving inference performance of the first model.

[00163] In some possible implementations, the N first modules further include N2 first modules, the first model includes N2 layers, each of the N2 layers includes one of the N2 first modules, $N2 \geq 2$, and N2 is a positive integer less than or equal to N.

[00164] The N first modules may include N2 first modules. In this way, the first active parameter may include an active parameter of the N2 first modules. $N2 \geq 2$, that is, the first active parameter may include active parameters of a plurality of first modules. The plurality of first modules may be located at different layers. Therefore, the first active parameter may include active parameters of a plurality of layers. To reduce a transmission latency, the second apparatus may deliver the active parameters of the plurality of layers at a time, a quantity of times of delivering the active parameters is reduced, and signaling overheads caused by frequent delivery of the active parameters are reduced. Optionally, N may be an integer from 1 to 8.

[00165] It should be noted that, the first model may include N2 layers, but is not limited to the N2 layers. For example, the first model may include more than N2 layers. Different from other layers in the first model, the N2 layers in the first model may be in one-to-one correspondence with the N2 first modules. For example, an mth layer in the N2 layers in the first model may include an mth first module of the N2 first modules, where m is an integer from 1 to N2. In this way, the first active parameter delivered by the second apparatus may include the active parameter of the N2 first modules at the N2 layers. A layer other than the N2 layers in the first model may include the first module, or may not include the first module. In other words, the first active parameter delivered by the second apparatus may include an active parameter of a module at a layer other than the N2 layers in the first model, or may not include the active parameter of the module at the layer other than the N2 layers in the first model.

[00166] Locations of the N2 layers of all layers in the first model are not limited in this application. For example, the N2 layers may be first N2 layers in the first model. In other words, the N2 layers may include a first layer to an $N2^{th}$ layer in the first model. For another example, the N2 layers may be intermediate N2 layers in the first model. In other words, the N2 layers may include an xth layer to an $(x+N2-1)^{th}$ layer in the first model, where x may be an integer greater than 1. In the foregoing two examples, numbers of the N2 layers of all layers are consecutive in the first model. In other words, no other layer is inserted into the N2 layers (that is, the N2 layers may include the layer of the first module, or may not include the layer of the first module). The first module may be understood as a module to which the active parameter delivered by the second

apparatus belongs. However, this is not limited in this application. Numbers of the N2 layers of all layers may be non-consecutive in the first model. In other words, another layer may be inserted into the N2 layers.

[00167] According to the foregoing embodiment, the second apparatus may deliver the active parameters of the plurality of layers at a time, a quantity of times of delivering the active parameters is reduced, and signaling overheads caused by frequent delivery of the active parameters are reduced.

[00168] With reference to the method 500 shown in FIG. 5, the N first modules may include the N1 first modules and the N2 first modules. In this way, the first active parameter may include active parameters of the N1 first modules and the N2 first modules. It should be noted that, in some possible implementations, the first information may be transmitted twice. In one transmission, the first information indicates a part of active parameters included in the first active parameter, for example, the active parameter of the N1 first modules. In another transmission, the first information indicates another part of active parameters included in the first active parameter, for example, the active parameter of the N2 first modules. However, the foregoing is merely an example, and is not intended to limit this application. For example, in some other possible implementations, the first information may be transmitted only once. The first active parameter indicated by the first information may include the active parameters of the N1 first modules and the N2 first modules. For another example, in some other possible implementations, the first information may be transmitted for more times. The first active parameter indicated by the first information transmitted each time may include a smaller quantity of active parameters than the first active parameter respectively indicated by the first information transmitted twice.

[00169] FIG. 6 is a schematic flowchart of still another communication method 600 according to an embodiment of this application. The method 600 may be combined with the method 400 and the method 500. In the method 600, full parameters of a second module at N2 layers in a first model may be stored locally in a first apparatus, and full parameters of a first module at the N2 layers in the first model may be stored in a second apparatus. In other words, the first model needs to request an active parameter of the first module at the N2 layers from the second apparatus. The second apparatus may determine an active parameter (that is, a first active parameter) based on the full parameters of the first module at the N2 layers. The following describes the method 600 with reference to FIG. 6.

[00170] In the solution shown in FIG. 6, it is assumed that the first module includes the 2nd one of a plurality of modules at each of a plurality of layers in the first model. It is assumed that the second module includes the 1st one of the plurality of modules at each of the plurality of layers in the first module. However, it should be noted that the description in FIG. 6 is merely used as an

example, and does not constitute a limitation on this application. For example, the first module may include the first and the 2nd one of the plurality of modules at each of the plurality of layers in the first model. In other words, in some other embodiments of this application, the first module may include all modules at the plurality of layers in the first module.

[00171] S610: The first apparatus predicts information about a location of first active parameter based on an input value of a layer preceding a 1st layer in the N2 layers. It should be noted that, in FIG. 6, an operation of S610 is shown by using a dashed line, indicating that S610 is an optional operation when S610 is combined with the method 400 and the method 500.

[00172] It may be understood that, before the N2 layers, the first model further includes another layer. For example, the first model includes 16 layers, and the N2 layers are a second layer to a ninth layer of the 16 layers. For example, a predictor module may predict the information about the location of the first active parameter based on the input value of the layer preceding the 1st layer in the N2 layers. Details are not described again. Refer to related descriptions of the method 400 and the method 500.

[00173] The input value of the layer preceding the 1st layer in the N2 layers may be understood as an input value of a 1st one of a plurality of modules at the layer preceding the 1st layer in the N2 layers.

[00174] In some possible implementations, when S610 is performed, the layer preceding the 1st layer in the N2 layers may perform inference based on the input value of the previous layer. In this way, when the layer preceding the 1st layer in the N2 layers performs inference, the predictor module may predict the information about the location of the first active parameter, that is, predict an active parameter of N2 first modules at the N2 layers after the layer that is performing inference. A layer for which an active parameter is predicted is not synchronized with a layer at which model inference is located. Therefore, the foregoing solution may also be referred to as asynchronous prediction.

[00175] A person skilled in the art may understand that, due to residual connections (residual connections), an output of each layer in a large model with a transformer structure changes very slowly. That is, outputs of adjacent layers are highly similar, so that the foregoing asynchronous multi-layer prediction solution almost does not reduce accuracy of predicting a location of an active parameter. In a possible implementation, N2 may be 8. That is, the first apparatus predicts, based on an input of one layer, location information of an active parameter of a first module at eight layers below the layer.

[00176] According to the foregoing embodiment, in one aspect, the first apparatus may predict information about locations of active parameters of a plurality of first modules that are respectively included in a plurality of layers. Compared with a solution in which an active parameter of only

one module is predicted at a time, the foregoing embodiment can save computing resources. In another aspect, the first apparatus predicts the N2 layers based on the input value of the layer preceding the 1st layer in the N2 layers, instead of based on an input value of the 1st layer in the N2 layers. In this way, the first apparatus may predict information about a location of an active parameter of a first module at the N2 layers when inferring the layer preceding the 1st layer in the N2 layers. In other words, a layer for which the first apparatus predicts the location of the active parameter is not a layer that is performing inference. Therefore, the foregoing embodiment does not affect the layer that is performing inference, that is, does not reduce inference performance of the large model.

[00177] In some other possible implementations, S610 may be replaced with the following: The first apparatus predicts the information about the location of the first active parameter based on the input value of the 1st layer in the N2 layers. A person skilled in the art may understand that the first apparatus obtains the input value of the 1st layer in the N2 layers, indicating that model inference is being performed at the 1st layer in the N2 layers. In this case, the predictor module is predicting the location of the first active parameter, that is, predicting an active parameter of the N2 first modules at the layer that is performing inference and (N2–1) layers after the layer. A layer for which an active parameter is predicted is synchronized with a layer at which model inference is located. Therefore, the foregoing solution may also be referred to as synchronous prediction. In another understanding, because the foregoing solution also predicts the layer after the layer at which model inference is located, the foregoing solution may also be referred to as asynchronous prediction.

[00178] With reference to the method 400, in some possible implementations, S410 is performed after S610. For example, the location of the first active parameter that is indicated by indication information included in third information may be predicted by the first apparatus based on the input value of the layer preceding the 1st layer in the N2 layers (or the input value of the 1st layer in the N2 layers). With reference to the method 400, in some possible implementations, S430 may be performed after S610.

[00179] Optionally, each of the N2 layers further includes one second module, and the second module is one of N2 second modules in the first model. In other words, each of the N2 layers may include one first module and one second module. For example, the second module may be an MHA module.

[00180] In some implementations, the method 600 further includes: (S620): The first apparatus predicts information about a location of a third active parameter based on the input value of the layer preceding the 1st layer in the N2 layers.

[00181] Optionally, the third active parameter is an active parameter of the N2 second modules.

For example, the third active parameter is an active parameter of N2 MHA modules at the N2 layers. A name of the third active parameter is not limited in this application. The third active parameter may also be referred to as a parameter, an active neuron, an active weight, an active element, or another name.

[00182] For example, the predictor module may predict the information about the location of the third active parameter based on the input value of the layer preceding the 1st layer in the N2 layers. Details are not described again. Refer to related descriptions of the method 400 and the method 500.

[00183] In some possible implementations, S610 and S620 may be simultaneously performed. In other words, the predictor module may predict the information about the location of the first active parameter and the information about the location of the third active parameter based on the input value of the layer preceding the 1st layer in the N2 layers.

[00184] In some possible implementations, when S620 is performed, the layer preceding the 1st layer in the N2 layers may perform inference based on the input value of the previous layer. In this way, when the layer preceding the 1st layer in the N2 layers performs inference, the predictor module may predict the information about the location of the third active parameter, that is, predict an active parameter of N2 second modules at the N2 layers after the layer that is performing inference. A layer for which an active parameter is predicted is not synchronized with a layer at which model inference is located. Therefore, the foregoing solution may also be referred to as asynchronous prediction.

[00185] According to the foregoing embodiment, in one aspect, the first apparatus may predict information about locations of active parameters of a plurality of second modules that are respectively included in a plurality of layers. Compared with a solution in which an active parameter of only one module is predicted at a time, the foregoing embodiment can save computing resources. In another aspect, a layer for which the first apparatus predicts the location of the active parameter is not a layer that is performing inference. Therefore, the foregoing embodiment does not affect the layer that is performing inference, that is, does not reduce inference performance of the large model.

[00186] In some other possible implementations, S620 may be replaced with the following: The first apparatus predicts the information about the location of the third active parameter based on the input value of the 1st layer in the N2 layers. A person skilled in the art may understand that the first apparatus obtains the input value of the 1st layer in the N2 layers, indicating that model inference is being performed at the 1st layer in the N2 layers. In this case, the predictor module is predicting the location of the third active parameter, that is, predicting an active parameter of the N2 second modules at the layer that is performing inference and N2–1 layers after the layer. A

layer for which an active parameter is predicted is synchronized with a layer at which model inference is located. Therefore, the foregoing solution may also be referred to as synchronous prediction. In another understanding, because the foregoing solution also predicts the layer after the layer at which model inference is located, the foregoing solution may also be referred to as asynchronous prediction.

[00187] In some possible implementations, the method 600 further includes: (S630): The first apparatus determines the third active parameter based on the information about the location of the third active parameter. For example, the first apparatus may store full parameters of the N2 second modules (for example, the MHA module) at the N2 layers in the first model. In this way, the first apparatus may determine a value of the third active parameter based on the information about the location of the third active parameter and the stored full parameters. In another possible implementation, the first apparatus sends request information to the second apparatus; and the second apparatus receives the request information. The request information may be used to request the third active parameter. In this case, the third active parameter may be obtained from the second apparatus. The second apparatus may prestore the full parameters of the second modules (for example, the MHA module) at the N2 layers in the first model, to determine the requested active parameter from the stored full parameters based on the request information of the first apparatus.

[00188] In some possible implementations, S630 may be performed before S610. For example, the execution sequence may be S620, S630, and S610. In some other possible implementations, S630 may be performed after S610. For example, the execution sequence may be S610, S620, and S630. For another example, the execution sequence may be S620, S610, and S630. However, this is not limited in this application. For example, S630 and S610 may be performed simultaneously.

[00189] In some possible implementations, the method 600 further includes: (S640): The first apparatus performs inference on the N2 layers according to the first active parameter and the third active parameter.

[00190] In some possible implementations, the performing inference on the N2 layers includes: The first apparatus uses an output value of a first module at a kth layer in the N2 layers as an input value of a second module at the kth layer. The foregoing solution may also be referred to as sequential inference.

[00191] In some possible implementations, the performing inference on the N2 layers includes: The first apparatus determines a third input value based on an output value of the first module at the kth layer in the N2 layers and an output value of the second module at the kth layer, where k is a positive integer less than or equal to N2. The foregoing solution may also be referred to as parallel inference.

[00192] For example, the first apparatus may add the output value of the first module at the kth

layer and the output value of the second module at the kth layer, to obtain the third input value. For another example, the first apparatus may add the output value of the first module at the kth layer and the output value of the second module at the kth layer, and multiply a value obtained after the addition by a non-zero coefficient, to obtain the third input value.

[00193] Optionally, the third input value is used as an input of a first module at a layer following the kth layer and an input of a second module at the layer following the kth layer.

[00194] FIG. 7 is a diagram of yet another communication method according to an embodiment of this application. FIG. 7 shows an embodiment of a combination of synchronous prediction and asynchronous prediction. In short, synchronous prediction is used for a first layer in a large model, and asynchronous prediction (or synchronous prediction) is used for a plurality of layers after the first layer. It should be noted that the embodiment shown in FIG. 7 is merely used as an example, and does not constitute a limitation on this application. The following provides descriptions with reference to FIG. 7.

[00195] FIG. 7 shows an inference process of first (2+2N2–1) layers of the large model. For ease of description, a first layer, a second layer to a $(2+N2-1)^{th}$ layer, and a $(2+N2)^{th}$ layer to a $(2+2N2-1)^{th}$ layer are respectively represented by dashed rounded rectangles. A number of a layer at which a module is located is represented by a subscript. For example, an MHA module at the first layer is represented by MHA_1. Other descriptions of an MHA module, an FNN module, and a predictor module are similar to related content in FIG. 3. Details are not described again.

[00196] Specifically, by performing embedding (embedding) on a request sent by a user to the large model, a generated vector or tensor may be used as a first "input" in FIG. 7. The first input may be an input to the MHA module at the first layer, as shown by MHA_1 and a dashed arrow pointing to MHA_1 in FIG. 7. The predictor module may predict a location of an active parameter of MHA_1 based on the input of MHA_1, which is represented by P1 and a solid arrow related to P1 in FIG. 7. A first apparatus may obtain the active parameter of MHA_1 from storage space for the first apparatus. The first apparatus may perform inference based on MHA_1 and the active parameter of MHA_1, to obtain an output of MHA_1. In other words, when MHA_1 performs inference, the first apparatus may load submatrices of a part of active heads to a memory of the first apparatus. The output of MHA_1 may be used as an input of the FNN module at the first layer, as shown by FNN_1 and a dotted arrow pointing to FNN_1 in FIG. 7. The predictor module may predict a location of an active parameter of FNN_1 based on the input of FNN_1, which is represented by P2 and a solid arrow related to P2 in FIG. 7. The first apparatus may request a second apparatus to deliver the active parameter of FNN_1. The second apparatus may store full parameters of the FNN module. For example, the second apparatus may store values of all elements in a 4×4 weight matrix and locations corresponding to each element. The second apparatus may determine a value of an active

parameter based on a location of an active parameter. For example, the location of the active parameter is a first row and a third column, and a second row and a fourth column. The second apparatus may determine, in the 4×4 weight matrix, values of elements in the first row and the third column and values of elements in the second row and the fourth column. The second apparatus may deliver the active parameter to the first apparatus. For example, the second apparatus may send first information to the first apparatus, where the first information indicates the active parameter of FNN_1 (that is, N1 first modules). For another example, the second apparatus may send, to the first apparatus, information used to indicate the values of elements in the first row and the third column and the second row and the fourth column. A process in which the first apparatus requests the active parameter and the second apparatus delivers the active parameter is shown by dashed double-headed arrows in FIG. 7. The first apparatus may perform inference based on FNN_1 and the active parameter of FNN_1. In other words, when FNN_1 performs inference, the first apparatus may load a part of weight parameters to the memory of the first apparatus.

[00197] A person skilled in the art may understand that a difference between two inputs of the first layer in the large model with the transformer structure is large. In the foregoing solution, synchronous prediction is performed on a first transformer layer, the location of the active parameter of MHA_1 is predicted by using the input of MHA_1, and the location of the active parameter of FNN_1 is predicted by using the input of FNN_1. In one aspect, the foregoing solution can save memory of the first apparatus. In another aspect, the active parameter at the first layer predicted in the foregoing solution is more accurate.

[00198] An output of FNN_1 may be used as an input of an MHA module at a second layer, as shown by MHA_2 and a dashed arrow pointing to MHA_2 in FIG. 7. The predictor module may predict active parameters of MHA modules and FNN modules at the second layer to the $(2+N2-1)^{th}$ layer based on the input of MHA_2, and are represented by P3 and a solid arrow related to P3 in FIG. 7. For the MHA module, for example, MHA_2 to MHA_{2+N2-1}, the first apparatus may obtain the active parameters of the MHA modules from storage space for the first apparatus. For the FNN module, for example, FNN_2 to FNN_{2+N2-1}, the first apparatus may request the second apparatus to deliver the active parameters of FNN_2 to FNN_{2+N2-1}. The second apparatus may deliver the active parameter to the first apparatus. For example, the second apparatus may send first information to the first apparatus, where the first information indicates the active parameters of FNN_2 to FNN_{2+N2-1} (that is, N2 first modules). For descriptions of MHA modules and FNN modules at the $(2+N2)^{th}$ layer to the $(2+2N2-1)^{th}$ layer, refer to the foregoing descriptions of the second layer to the $(2+N2-1)^{th}$ layer. Details are not described again.

[00199] A person skilled in the art may understand that, outputs of the large model with the

transformer structure at the second layer and layers after the second layer are highly similar. Therefore, an input of a layer may be used to predict active parameters of a plurality of subsequent layers at a time. In this way, the second apparatus may deliver active parameters of a plurality of layers at a time, to effectively reduce overheads caused by information transmission between the first apparatus and the second apparatus.

[00200] FIG. 8A and FIG. 8B are a schematic flowchart of yet another communication method 800 according to an embodiment of this application. In FIG. 8A and FIG. 8B, a first apparatus may include a user operation module, a large model inference user interface (user interface, UI), a predictor module, and a terminal-side inference module. Optionally, the user operation module and the large model inference UI may also be apparatuses that match the first apparatus. The method 800 may be combined with the procedures of the first apparatus and the second apparatus shown in FIG. 7. It should be noted that the method 800 is merely used as an example, and does not constitute a limitation on this application. The following describes the method 800.

[00201] S802: The user operation module sends request information to the large model inference user interface (user interface, UI). In other words, the user operation module may send a request to the large model inference UI. The user operation module may convert a user operation into information that can be identified by the first apparatus. Optionally, the request information includes an input (input) of a user.

[00202] S804: The large model inference UI sends the input of the user to the predictor module, and the predictor module receives the input of the user. Before the predictor module performs prediction, the first apparatus may perform processing such as embedding (embedding) on the input of the user, and an obtained vector or tensor may be used as an input of MHA_1 at a first layer.

[00203] S810: The predictor module predicts a location of an active parameter of MHA_1 based on the input of MHA_1.

[00204] S812: The predictor module may send indication information of the location of the active parameter of MHA_1 to the terminal-side inference module, and the terminal-side inference module receives the indication information of the location of the active parameter of MHA_1.

[00205] S814: The terminal-side inference module performs inference based on the active parameter of MHA_1, the input of MHA_1, and MHA_1, to obtain an output of MHA_1. The terminal-side inference module may first obtain the active parameter of MHA_1 from storage space for the first apparatus based on the location of the active parameter of MHA_1, and then perform inference based on the active parameter of MHA_1, the input of MHA_1, and MHA_1.

[00206] S816: The terminal-side inference module sends the output of MHA_1 (that is, an input of FNN_1) to the predictor module, and the predictor module receives the output of MHA_1 (that is, the input of FNN_1).

[00207] S818: The predictor module predicts a location of an active parameter of FNN_1 based on the input of FNN_1.

[00208] S820: The predictor module may send, to a second apparatus, information used to request the active parameter of FNN_1, and the second apparatus may receive the information used to request the active parameter of FNN_1. The information used to request the active parameter of FNN_1 may include indication information of the location of the active parameter of FNN_1. It should be noted that the method 800 is merely for ease of understanding, and the operation of sending the request to the second apparatus may also be performed by another module, for example, a sending module.

[00209] S822: The second apparatus sends the active parameter of FNN_1 to the first apparatus, and the first apparatus receives the active parameter of FNN_1. In some possible implementations, the terminal-side inference module in the first apparatus may receive the active parameter of FNN_1. However, this is not limited in this application. The operation of receiving by the first apparatus may also be performed by another module, for example, a receiving module.

[00210] S824: The terminal-side inference module performs inference based on the active parameter of FNN_1, the input of FNN_1, and FNN_1, to obtain an output of FNN_1.

[00211] S830: The terminal-side inference module sends the output of FNN_1 (that is, an input of MHA_2) to the predictor module, and the predictor module receives the output of FNN_1 (that is, the input of MHA_2).

[00212] S832: The predictor module predicts locations of active parameters of MHA_2 to MHA_{2+N2-1} and locations of active parameters of FNN_2 to FNN_{2+N2-1} based on the input of MHA_2.

[00213] S834: The predictor module may send, to the second apparatus, information used to request the active parameters of FNN_2 to FNN_{2+N2-1}, and the second apparatus may receive the information used to request the active parameters of FNN_2 to FNN_{2+N2-1}. The information used to request the active parameters of FNN_2 to FNN_{2+N2-1} may include indication information of the locations of the active parameters of FNN_2 to FNN_{2+N2-1}. It should be noted that the method 800 is merely for ease of understanding, and the operation of sending the request to the second apparatus may also be performed by another module, for example, a sending module.

[00214] S836: The predictor module may send indication information of the locations of the active parameters of MHA_2 to MHA_{2+N2-1} to the terminal-side inference module, and the terminal-side inference module may receive the indication information of the locations of the active parameters of MHA_2 to MHA_{2+N2-1}.

[00215] S838: The second apparatus sends the active parameters of FNN_2 to FNN_{2+N2-1} to the first apparatus. In some possible implementations, the terminal-side inference module in the first apparatus may receive the active parameters of FNN_2 to FNN_{2+N2-1}. However, this is not limited

in this application. The operation of receiving by the first apparatus may also be performed by another module, for example, a receiving module.

[00216] It should be noted that an execution sequence of S834 to S838 is not limited. For example, the execution sequence may be S834, S838, and S836. For another example, the execution sequence may be S836, S834, and S838.

[00217] S840: The terminal-side inference module performs inference based on the modules at the second layer to the $(2+N2-1)^{th}$ layer, the active parameter of the module, and the input of the module, and finally outputs an output of FNN_{2+N2-1} (that is, an input of MHA_{2+N2}). For MHA_2 to MHA_{2+N2-1}, the terminal-side inference module may first respectively obtain the active parameters of MHA_2 to MHA_{2+N2-1} from storage space for the first apparatus based on the locations of the active parameters of MHA_2 to MHA_{2+N2-1}.

[00218] Then, the predictor module may predict active parameters of modules at subsequent N2 layers based on the input of MHA_{2+N2}, and perform operations similar to S832 to S840 until an inference result of a large model is output. The inference result of the large model can be returned to the large model inference UI.

[00219] In some possible implementations, the first model includes M layers, M is a positive integer greater than N2, and numbers of the N2 layers among the M layers are non-consecutive. In other words, the first model includes M layers with consecutive numbers, and the M layers include N2 layers.

[00220] That numbers of the N2 layers among the M layers are non-consecutive may be understood as that numbers of two adjacent layers among the N2 layers are non-consecutive. For example, the M layers are first 10 layers in the first model, and the N2 layers are respectively a second layer, a third layer, a fifth layer, and a sixth layer. The third layer is adjacent to the fifth layer, but the number 3 and the number 5 are not consecutive.

[00221] A person skilled in the art may understand that some transformer layers are randomly skipped, for example, 25% of all layers included in a large model, so that performance of the large model is not greatly reduced, and an inference latency can be reduced. The large model in this embodiment of this application supports inference of randomly skipping one layer. For example, one transformer layer may be skipped for every four transformer layers. For example, the N2 layers are respectively the second layer, the third layer, the fifth layer, and the sixth layer. An output of the second layer may be used as an input of the third layer, and an output of the third layer may be used as an input of the fifth layer. In other words, a fourth layer is skipped and does not participate in inference. For another example, N2 is 8, and a fourth layer and an eighth layer may be skipped in an inference process of the large model.

[00222] Correspondingly, in the solution provided in this embodiment of this application, when

a plurality of layers after a current inference layer are predicted, some layers may be skipped, that is, active parameters of these layers are not predicted. With reference to the description of S410, the first information delivered by the second apparatus may indicate the first active parameter, that is, indicate the active parameter of the N first modules. The N first modules may include N2 first modules, and the N2 first modules respectively belong to N2 layers. In this way, the first active parameter may include the active parameter of the first module at the N2 layers. Numbers of the N2 layers among the M layers are not consecutive, indicating that the active parameter indicated by the second apparatus does not belong to the N2 layers with consecutive numbers, but belongs to the N2 layers with non-consecutive numbers, that is, one or more layers may be skipped for the active parameter indicated by the second apparatus. With reference to the description of S610, the first active parameter predicted by the first apparatus does not belong to active parameters of the N2 layers with consecutive numbers, but belongs to active parameters of the N2 layers with non-consecutive numbers, that is, the first apparatus skips some layers when predicting the active parameter, to reduce a prediction latency.

[00223] According to the foregoing embodiment, the first active parameter indicated by the first information may include active parameters of layers with non-consecutive numbers, that is, one or more layers may be skipped for these active parameters, to reduce an amount of information for transmitting the first information and reduce latency.

[00224] In some possible implementations, N2 is related to a resource for transmitting the first information and/or an inference latency requirement.

[00225] Optionally, the resource for transmitting the first information may include a bandwidth, a spectral width, a timeslot, a multi-antenna beam, and the like that are used to transmit the first information. For example, a larger bandwidth used to transmit the first information may indicate a larger N2. On the contrary, a smaller bandwidth used to transmit the first information indicates a smaller N2.

[00226] The inference latency requirement may be a latency requirement for performing inference by the first apparatus. For example, the first apparatus needs to complete inference of one layer within 3 seconds (s). A higher inference latency requirement, or a smaller required inference latency indicates a larger N2. On the contrary, a lower inference latency requirement, or a larger required inference latency may indicate a smaller N2.

[00227] Optionally, the inference latency requirement may include a latency upper limit of obtaining the first information by the first apparatus. For example, if the first apparatus needs to complete inference of one layer within 3s, and a time spent by the first apparatus in performing inference based on the active parameter is 2s, the first apparatus obtains a latency upper limit of the first information used to indicate an active parameter of the N2 first modules (including the

layer and the (N2−1) layers after the layer) is 1s, and a latency of obtaining the first information by the first apparatus cannot exceed 1s. Correspondingly, a latency of sending the first information by the second apparatus cannot exceed 1s. In other words, the first apparatus needs to obtain the first information within 1s, to obtain the active parameter of the N2 first modules.

[00228] Using GPT3 as an example, GPT3 has 96 transformer layers in total, and storage space occupied by an active parameter of each transformer layer is 175 B×0.15/96×4=0.09375 GB=93.75 MB. Storage space occupied by the active parameter of the FNN module is 93.75×2/3=62.5 MB. If a bandwidth is 200 megabytes per second (MB/s), and the latency upper limit of receiving the first information by the first apparatus is 1s, N2 may be 3. In this way, transmission of the first information can meet the inference latency requirement within given resources.

[00229] According to the foregoing embodiment, a value of N2 may be related to the resource for transmitting the first information and/or the inference latency requirement, so that transmission of the first information can meet a resource limitation and/or the inference latency requirement. The foregoing solution can ensure that the first information can be successfully transmitted, or can ensure that transmission of the first information can meet the inference latency requirement of the first model, to further ensure inference performance of the first model.

[00230] FIG. 9 is a diagram of a large model used for parallel inference according to an embodiment of this application. It should be noted that FIG. 9 is merely used as an example, and does not constitute a limitation on this application.

[00231] Refer to FIG. 9. An input of a first layer may be used as an input of MHA_1 and an input of FNN_1. An output of the first layer may be determined based on an output of MHA_2 and an output of FNN_2, that is, the output of the first layer is used as an input of a second layer. The inference manner shown in FIG. 9 may be referred to as parallel inference. It should be noted that k in S650 is a positive integer less than or equal to N2. In some possible implementations, k is a positive integer from 1 to N2. In other words, the foregoing parallel inference manner is used for each of the N2 layers. In some other possible implementations, k is a part of the positive integers from 1 to N2. In other words, a part of the N2 layers (or one or more layers) use the foregoing parallel inference manner, and remaining part of layers still use an inference manner similar to that in FIG. 3 and FIG. 7. The inference manner may be referred to as sequential inference. That is, an input of a layer is used as an input of an MHA module at the layer, an output of the MHA module is used as an input of an FNN module at the layer, and an output of the FNN module is used as an output of the layer.

[00232] Compared with the sequential inference manner, the parallel inference manner may reduce a quantity of inputs at each layer by half. For example, when N2 is 10, if the 10 layers all

use the sequential inference manner, each layer has two inputs, and the 10 layers have a total of 20 inputs. If the 10 layers all use the parallel inference manner, each layer has only one input, and the 10 layers have a total of 10 inputs.

[00233] A quantity of inputs at each layer is reduced by half, and a change rate of adjacent inputs (for example, inputs of adjacent modules in sequential inference, or inputs of adjacent layers in parallel inference) is fixed. Therefore, compared with the sequential inference manner, the parallel inference manner may enable a predictor module to predict locations of active parameters of twice the number of layers. For example, in the sequential inference manner, when N2 may be 8, that is, in a case in which locations of eight layers of active parameters are predicted at a time, locations of predicted active parameters are accurate. In this case, in the parallel inference manner, N2 may be 16, that is, locations of 16 layers of active parameters can be predicted at a time (it is assumed that parallel inference is used for all the 16 layers). In addition, the locations of the eight layers of active parameters can be as accurate as those predicted in the case of sequential inference.

[00234] According to the foregoing embodiment, one or more layers among the N2 layers may use a form of parallel inference. That is, when a first module at a layer performs inference, a second module at the layer may also perform inference. Sequential inference is to use an output obtained after inference of a second module at a layer is completed as an input of the second module at the layer. Compared with the layer using sequential inference, the layer using parallel inference can reduce a quantity of inputs from 2 to 1. A quantity of inputs is reduced, and a change rate of adjacent inputs is fixed, so that a value of N2 that can be selected may be increased, and the active parameter indicated by the first information delivered by the second apparatus may be active parameters of more layers. In other words, the second apparatus may deliver the active parameters of more layers at a time, to further reduce a transmission latency from the second apparatus to the first apparatus.

[00235] With reference to some modules in FIG. 8A and FIG. 8B, in some possible implementations, a terminal-side inference module may determine confidence of an output value of a first model, and send instructions to the predictor module when the confidence of the output value of the first model is greater than or equal to a first threshold, where the instructions are used to instruct the predictor module to stop prediction. In some possible implementations, when the confidence of the output value of the first model is greater than or equal to the first threshold, the first model may stop inference. In some possible implementations, when the confidence of the output value of the first model is greater than or equal to the first threshold, the first model may send the output value of the first model (or referred to as an inference result of the first model) to a large model inference UI and/or a user operation module.

[00236] According to the foregoing embodiment, the first apparatus may send the second

information when the confidence of the output value of the first model meets a condition, where the second information may indicate stopping transmission of the first information, to reduce a resource waste caused by transmitting the first information. In addition, the foregoing embodiment may be applied to an inference early exiting scenario (early exiting). In the scenario in which the first model exits inference early, the first model exits inference not because inference is completed at all layers in the first model, but because the confidence of the output value of the first model meets the condition. The early exit of inference may enable a part of layers in the first model not to participate in inference, to reduce an inference latency. In the foregoing embodiment, when the first model exits the inference early, the second apparatus can be prevented from transmitting information indicating an active parameter used for inference, to save transmission resources.

[00237] The following describes apparatus embodiments corresponding to the method embodiments of this application. The following merely briefly describes apparatuses. For specific implementation steps and details of the solutions, refer to the foregoing method embodiments.

[00238] To implement functions in the method provided in this application, a communication apparatus may include a hardware structure and/or a software module, and the foregoing functions are implemented in a form of a hardware structure, a software module, or a combination of a hardware structure and a software module. Whether a function in the foregoing functions is performed by the hardware structure, the software module, or the combination of the hardware structure and the software module depends on particular applications and design constraints of the technical solutions.

[00239] FIG. 10 is a block diagram of a communication apparatus 1000 according to an embodiment of this application. The communication apparatus 1000 includes a processor 1010 and a communication interface 1020. Optionally, the processor 1010 and the communication interface 1020 may be connected to each other through a bus. The communication apparatus 1000 may be a first apparatus, or may be a second apparatus. For example, the first apparatus may be a terminal device or a terminal-side device, and the second apparatus may be a network device, a core network device, a cloud-side device, or an edge-side device.

[00240] Optionally, the communication apparatus 1000 may further include a memory 1040. The memory 1040 includes but is not limited to a random access memory (random access memory, RAM), a read-only memory (read-only memory, ROM), a cache (cache), an erasable programmable read-only memory (erasable programmable read-only memory, EPROM), a synchronous dynamic random access memory (synchronous dynamic random access memory, SDRAM), a hard disk drive (hard disk drive, HDD), a solid-state drive (solid-state drive, SSD), or a portable read-only memory (compact disc read-only memory, CD-ROM). The memory 1040 is configured to store related instructions and/or data. The memory 1040 and the processor 1010 may

be integrated together or separately disposed.

[00241] For example, the communication apparatus 1000 may be the first apparatus. Memory of the first apparatus may include a RAM, a cache, an SDRAM, or a CD-ROM. Optionally, the memory of the first apparatus may be configured to load a first active parameter. Flash memory of the first apparatus may include a ROM, an EPROM, an HDD, or an SSD. Optionally, the flash memory of the first apparatus may store full parameters of a second module.

[00242] For example, the communication apparatus 1000 may be the second apparatus. Flash memory of the second apparatus may include a ROM, an EPROM, an HDD, or an SSD. Optionally, the flash memory of the first apparatus may store full parameters of a first module.

[00243] The processor 1010 may be one or more central processing units (central processing units, CPUs). When the processor 1010 is one CPU, the CPU may be a single-core CPU, or may be a multi-core CPU. However, this is not limited in this application. The processor 1010 may be one or more GPUs, or may be one or more TPUs. The processor 1010 may be a signal processor, a chip, or another integrated circuit that can implement the methods in this application, or may be a part of circuits configured to perform a processing function in the foregoing processor, chip, or integrated circuit. In addition, the communication interface 1020 may be an input/output interface, where the input/output interface is configured to input or output a signal or data, or may be an input/output circuit.

[00244] For example, the communication apparatus 1000 is the first apparatus, and the processor 1010 is configured to perform the following operations: receiving first information, where the first information indicates a first active parameter, the first active parameter is an active parameter of N first modules in a first model, and N is a positive integer; and performing inference based on the N first modules according to the first active parameter.

[00245] For example, the communication apparatus 1000 is the second apparatus, and the processor 1010 is configured to perform the following operations: generating first information, where the first information indicates a first active parameter, the first active parameter is an active parameter of N first modules in a first model, the first active parameter is used to perform inference based on the N first modules, and N is a positive integer; and sending the first information.

[00246] The foregoing content is merely used as an example for description. The communication apparatus 1000 is responsible for performing related methods or steps of the first apparatus or the second apparatus in the foregoing method embodiments.

[00247] In a possible implementation, the communication interface 1020 may be a transceiver. The transceiver may include a transmitter and a receiver. The transmitter is configured to perform a sending operation, and the receiver is configured to perform a receiving operation. For example, the processor 1010 is configured to control the transceiver to receive and/or send a signal.

[00248] In a possible implementation, the communication interface 1020 may alternatively be a communication circuit, a pin, an input/output interface, a bus, or the like.

[00249] It should be noted that the communication apparatus 1000 may include the transmitter, but does not include the receiver. Alternatively, the communication apparatus 1000 may include the receiver, but does not include the transmitter. This may specifically depend on whether the foregoing solution performed by the communication apparatus 1000 includes a sending action and a receiving action.

[00250] The foregoing descriptions are merely examples. For specific content, refer to the content shown in the method embodiments. For implementations of the operations in FIG. 10, further correspondingly refer to the corresponding descriptions of the method embodiments shown in FIG. 4 to FIG. 9.

[00251] For example, the communication apparatus 1000 may be configured to perform the solutions shown in FIG. 4 to FIG. 9.

[00252] For example, the communication apparatus 1000 is the first apparatus, and the communication interface 1020 may be configured to receive first information.

[00253] For example, the communication apparatus 1000 is the second apparatus, and the communication interface 1020 may be configured to send the first information.

[00254] For details of other implementations, refer to the detailed descriptions of the embodiments shown in FIG. 4 to FIG. 9. Details are not described herein again. It should be understood that a specific process in which the components perform the foregoing corresponding processes is described in detail in the foregoing method embodiments. For brevity, details are not described herein again.

[00255] FIG. 11 is a block diagram of another communication apparatus 1100 according to an embodiment of this application. The communication apparatus 1100 may be a first apparatus or a second apparatus, or may be a chip or a module in the first apparatus or the second apparatus, and is configured to implement the methods in the embodiments shown in FIG. 4 to FIG. 9. For details, refer to related descriptions in the foregoing method embodiments.

[00256] The communication apparatus 1100 includes a transceiver unit 1110. The following describes the transceiver unit 1110 for example.

[00257] The transceiver unit 1110 may include a sending unit and a receiving unit. The sending unit is configured to perform a sending action of the communication apparatus, and the receiving unit is configured to perform a receiving action of the communication apparatus. For ease of description, in embodiments of this application, the sending unit and the receiving unit are combined into one transceiver unit. Unified descriptions are provided herein, and details are not described below again. The transceiver unit 1110 may implement a corresponding communication

function. The transceiver unit 1110 may also be referred to as a communication interface or a communication module.

[00258] It should be noted that the communication apparatus 1100 may include the sending unit, but does not include the receiving unit. Alternatively, the communication apparatus 1100 may include the receiving unit, but does not include the sending unit. This may specifically depend on whether the foregoing solution performed by the communication apparatus 1100 includes a sending action and a receiving action.

[00259] For example, the transceiver unit 1110 is configured to receive first information and the like.

[00260] Optionally, the communication apparatus 1100 may further include a processing unit 1120 configured to perform content related to steps such as processing and coordination of the communication apparatus 1100. For example, the processing unit 1120 is configured to perform inference based on N first modules according to first active parameter.

[00261] For example, the transceiver unit 1110 is configured to send the first information and the like.

[00262] Optionally, the communication apparatus 1100 may further include a processing unit 1120 configured to perform content related to steps such as processing and coordination of the communication apparatus 1100. For example, the processing unit 1120 is configured to generate first information.

[00263] The foregoing content is merely used as an example for description. The communication apparatus 1100 is responsible for performing related methods or steps in the foregoing method embodiments.

[00264] Optionally, the communication apparatus 1100 further includes a storage unit 1130. The storage unit 1130 is configured to store a program or code used to perform the foregoing methods. In other words, the storage unit 1130 may be configured to store instructions and/or data, and the processing unit 1120 may read the instructions and/or the data in the storage unit 1130, to enable the communication apparatus 1100 to implement the foregoing method embodiments. For example, the communication apparatus 1100 may be configured to perform the solutions shown in FIG. 4 to FIG. 9.

[00265] For example, the transceiver unit 1110 may be configured to receive first information, where the first information indicates a first active parameter, the first active parameter is an active parameter of N first modules in a first model, and N is a positive integer; and the processing unit 1120 may be configured to perform inference based on the N first modules according to the first active parameter.

[00266] For example, the processing unit 1120 may be configured to generate first information,

where the first information indicates a first active parameter, the first active parameter is an active parameter of N first modules in a first model, the first active parameter is used to perform inference based on the N first modules, and N is a positive integer; and the transceiver unit 1110 may be configured to send the first information.

[00267] For details of other implementations, refer to the detailed descriptions of the embodiments shown in FIG. 4 to FIG. 9. Details are not described herein again. It should be understood that a specific process in which the components perform the foregoing corresponding processes is described in detail in the foregoing method embodiments. For brevity, details are not described herein again.

[00268] When the communication apparatus 1100 in FIG. 10 is a chip, the communication interface 1120 may be a transceiver, an input/output circuit, or a communication interface of the chip. The processor 1110 may be a processor, a microprocessor, or an integrated circuit integrated on the chip. A sending operation of the first apparatus or the second apparatus in the foregoing method embodiments may be understood as an output of the chip, and a receiving operation of the first apparatus or the second apparatus in the foregoing method embodiments may be understood as an input of the chip.

[00269] When the communication apparatus 1100 in FIG. 10 is a chip, the transceiver unit 1110 may be a transceiver, an input/output circuit, or a communication interface of the chip. The processing unit 1120 may be a processor, a microprocessor, or an integrated circuit integrated on the chip. A sending operation of the first apparatus or the second apparatus in the foregoing method embodiments may be understood as an output of the chip, and a receiving operation of the first apparatus or the second apparatus in the foregoing method embodiments may be understood as an input of the chip.

[00270] This application further provides a chip, including a processor, configured to invoke instructions from a memory and run the instructions stored in the memory, to enable a communication apparatus on which the chip is installed to perform the method in the foregoing examples.

[00271] This application further provides another chip, including an input interface, an output interface, and a processor. The input interface, the output interface, and the processor are connected through an internal connection path. The processor is configured to execute code in a memory. When the code is executed, the processor is configured to perform the method in the foregoing examples. Optionally, the chip further includes the memory. The memory is configured to store a computer program or the code.

[00272] This application further provides a processor, configured to be coupled to a memory and configured to perform the method and function of the communication apparatus in any one of

the foregoing embodiments.

[00273] Another embodiment of this application provides a computer program product including a computer program or instructions. When the computer program product is run on a computer, the method in the foregoing embodiments is implemented.

[00274] This application further provides a computer program. When the computer program is run on a computer, the method in the foregoing embodiments is implemented.

[00275] Another embodiment of this application provides a computer-readable storage medium. The computer-readable storage medium stores a computer program. When the computer program is run on a computer, the method in the foregoing embodiments is implemented.

[00276] This application further provides a communication system. The communication system includes a first apparatus and a second apparatus. The first apparatus and the second apparatus are respectively configured to perform the methods performed by the first apparatus and the second apparatus in the foregoing embodiments.

[00277] A person of ordinary skill in the art may be aware that, in combination with the examples described in embodiments disclosed in this specification, units and algorithm steps may be implemented by electronic hardware or a combination of computer software and electronic hardware. Whether the functions are performed by hardware or software depends on particular applications and design constraint conditions of the technical solutions. A person skilled in the art may use different methods to implement the described functions for each particular application, but it should not be considered that the implementation goes beyond the scope of this application.

[00278] It may be clearly understood by a person skilled in the art that, for the purpose of convenient and brief description, for a detailed working process of the foregoing system, apparatus, and unit, refer to a corresponding process in the foregoing method embodiments. Details are not described herein again.

[00279] In the several embodiments provided in this application, it should be understood that the disclosed system, apparatus, and method may be implemented in other manners. For example, the described apparatus embodiments are merely examples. For example, the unit division is merely logical function division and may be other division in actual implementation. For example, a plurality of units or components may be combined or integrated into another system, or some features may be ignored or not performed. In addition, the displayed or discussed mutual couplings or direct couplings or communication connections may be implemented through some interfaces. The indirect couplings or communication connections between the apparatuses or units may be implemented in electronic, mechanical, or other forms.

[00280] The units described as separate components may or may not be physically separate, and components displayed as units may or may not be physical units, may be located in one

location, or may be distributed on a plurality of network units. A part or all of the units may be selected based on actual requirements to achieve the objectives of the solutions of embodiments.

[00281] In addition, functional units in embodiments of this application may be integrated into one processing unit, each of the units may exist independently physically, or two or more units are integrated into one unit.

[00282] When the functions are implemented in a form of a software functional unit and sold or used as an independent product, the functions may be stored in a computer-readable storage medium. Based on such an understanding, the technical solutions of this application essentially, the part contributing to the conventional technology, or a part of the technical solutions may be implemented in a form of a software product. The computer software product is stored in a storage medium, and includes several instructions for enabling a computing device (which may be a personal computer, a server, a network device, or the like) to perform all or a part of the steps of the methods described in embodiments of this application. The storage medium includes various media that can store program code, such as a USB flash drive, a portable hard disk drive, a read-only memory (read-only memory, ROM), a random access memory (random access memory, RAM), a magnetic disk, and an optical disk.

[00283] The foregoing descriptions are merely specific implementations of this application, but are not intended to limit the protection scope of this application. Any variation or replacement readily figured out by a person skilled in the art within the technical scope disclosed in this application shall fall within the protection scope of this application. Therefore, the protection scope of this application shall be subject to the protection scope of the claims.

CLAIMS

1. A communication method, wherein the method comprises:

receiving first information, wherein the first information indicates a first active parameter, the first active parameter is an active parameter of N first modules in a first model, and N is a positive integer; and

performing inference based on the N first modules according to the first active parameter.

2. The method according to claim 1, wherein the N first modules comprise N1 first modules, the N1 first modules are modules at a first layer of at least one layer in the first model, and N1 is a positive integer less than or equal to N.

3. The method according to claim 2, wherein the N1 first modules are a 2nd one of a plurality of modules at the first layer in the first model, N1=1, and the method further comprises:

predicting information about a location of the first active parameter based on a second input value, wherein the second input value is an input value of the 2nd one of the plurality of modules at the first layer in the first model or an output value of a 1st one of the plurality of modules at the first layer in the first model.

4. The method according to claim 2 or 3, wherein the method further comprises:

predicting information about a location of a second active parameter based on a first input value, wherein the second active parameter is an active parameter of the 1st one of the plurality of modules at the first layer in the first model, and the first input value is an input value of the 1st one of the plurality of modules at the first layer in the first model.

5. The method according to any one of claims 1 to 4, wherein the N first modules further comprise N2 first modules, the first model comprises N2 layers, each of the N2 layers comprises one of the N2 first modules, $N2 \geq 2$, and N2 is a positive integer less than or equal to N.

6. The method according to claim 5, wherein the first model comprises M layers, M is a positive integer greater than N2, and numbers of the N2 layers among the M layers are non-consecutive.

7. The method according to claim 5 or 6, wherein N2 is related to a resource for transmitting the first information and/or an inference latency requirement.

8. The method according to any one of claims 5 to 7, wherein the method further comprises:

predicting the information about the location of the first active parameter based on an input value of a layer preceding a 1st layer in the N2 layers.

9. The method according to any one of claims 5 to 8, wherein each of the N2 layers further comprises one second module, the second module is one of N2 second modules in the first model, and the method further comprises:

predicting information about a location of a third active parameter based on the input value of the layer preceding the 1st layer in the N2 layers, wherein the third active parameter is an active parameter of the N2 second modules.

10. The method according to claim 9, wherein the second module is a multi-head attention MHA module.

11. The method according to any one of claims 5 to 10, wherein each of the N2 layers further comprises one second module, and the method further comprises:

determining a third input value based on an output value of a first module at a kth layer in the N2 layers and an output value of a second module at the kth layer, wherein the third input value is used as an input of a first module at a layer following the kth layer and an input of a second module at the layer following of the kth layer, and k is a positive integer less than or equal to N2.

12. The method according to any one of claims 1 to 11, wherein the method further comprises:

sending second information when confidence of an output value of the first model is greater than or equal to a first threshold, wherein the second information indicates stopping transmission of the first information.

13. The method according to any one of claims 1 to 12, wherein the method further comprises:

sending third information, wherein the third information comprises the information indicating the location of the first active parameter, and the third information is used to request the first active parameter.

14. The method according to any one of claims 1 to 13, wherein the first module is a feedforward neural network FNN module.

15. A communication method, wherein the method comprises:

generating first information, wherein the first information indicates a first active parameter, the first active parameter is an active parameter of N first modules in a first model, the first active parameter is used to perform inference based on the N first modules, and N is a positive integer; and

sending the first information.

16. The method according to claim 15, wherein the N first modules comprise N1 first modules, the N1 first modules are modules at a first layer of at least one layer in the first model, and N1 is a positive integer less than or equal to N.

17. The method according to claim 15 or 16, wherein the N first modules further comprise N2 first modules, the first model comprises N2 layers, each of the N2 layers comprises one of the N2 first modules, $N2 \geq 2$, and N2 is a positive integer less than or equal to N.

18. The method according to claim 17, wherein the first model comprises M layers, M is a positive integer greater than N2, and numbers of the N2 layers among the M layers are non-

consecutive.

19. The method according to claim 17 or 18, wherein N2 is related to a resource for transmitting the first information and/or an inference latency requirement.

20. The method according to any one of claims 15 to 19, wherein the method further comprises:

receiving second information, wherein the second information indicates stopping transmission of the first information.

21. The method according to any one of claims 15 to 20, wherein the method further comprises:

receiving third information, wherein the third information comprises indication information of a location of the first active parameter.

22. The method according to claim 21, wherein the sending the first information comprises:

sending the first information in response to the third information.

23. The method according to any one of claims 15 to 22, wherein the first module is a feedforward neural network FNN module.

24. A communication apparatus, comprising at least one module or at least one unit, wherein the at least one module or the at least one unit is configured to perform the method according to any one of claims 1 to 15 or the method according to any one of claims 16 to 23.

25. A communication apparatus, comprising a processor, wherein the processor is configured to execute a computer program or instructions, to enable the communication apparatus to perform the method according to any one of claims 1 to 14 or the method according to any one of claims 15 to 23.

26. The communication apparatus according to claim 25, wherein the communication apparatus further comprises a memory, and the memory is configured to store the computer program or the instructions.

27. A computer-readable storage medium, wherein the computer-readable storage medium stores a computer program or instructions, and when the computer program or the instructions are run on a computer, the method according to any one of claims 1 to 14 or the method according to any one of claims 15 to 23 is performed.

28. A computer program product, comprising a computer program or instructions, wherein when the computer program or the instructions are run, the method according to any one of claims 1 to 14 or the method according to any one of claims 15 to 23 is implemented.

29. A communication system, comprising a first apparatus and a second apparatus, wherein the first apparatus is configured to perform the method according to any one of claims 1 to 14, and the second apparatus is configured to perform the method according to any one of claims 15 to 23.

30. A chip, comprising a processor, configured to invoke a computer program or computer instructions in a memory, to enable the chip to perform the method according to any one of claims 1 to 14 or the method according to any one of claims 15 to 23.