COMMUNICATION METHOD AND RELATED APPARATUS

[0001] This application claims priority to Chinese Patent Application No. 202410303028.6, filed with the China National Intellectual Property Administration on March 15, 2024 and entitled "COMMUNICATION METHOD AND RELATED APPARATUS", which is incorporated herein by reference in its entirety.

TECHNICAL FIELD

[0002] This application relates to the communication field, and in particular, to a communication method and a related apparatus.

BACKGROUND

[0003] Wireless communication may be transmission communication performed between two or more communication nodes without propagation relying on conductors or cables, and the communication nodes generally include a network device and a terminal device.

[0004] Currently, in a wireless communication system, a communication node usually has signal receiving and sending capabilities and a computational capability. A network device having a computational capability is used as an example. The computational capability of the network device is mainly used to provide computational power support for signal receiving and sending capabilities (for example, sending and receiving a signal), to implement communication between the network device and another communication node.

[0005] However, in a communication network, in addition to providing computational power support for the foregoing communication task based on the computational capability of the communication node, the communication node may further have spare computational capabilities. Therefore, how to use these computational capabilities is a technical problem.

SUMMARY

[0006] This application provides a communication method and a related apparatus to enhance robustness of continual learning.

[0007] A first aspect of this application provides a communication method. The method is applied to a first communication apparatus. The first communication apparatus may be a

communication device (for example, a terminal device or a network device), or the first communication apparatus may be some components (for example, a processor, a chip, or a system on chip) in the communication device, or the first communication apparatus may be a logic module or software that can implement all or some functions of the communication device. In the method, the first communication apparatus obtains first information, where the first information indicates performance monitoring information of at least one artificial intelligence (artificial intelligence, AI) model, and the at least one AI model is associated with a first AI model deployed in the first communication apparatus; and the first communication apparatus performs noise perturbation on a model parameter of the first AI model based on the first information, to obtain an updated model parameter of the first AI model.

[0008] Based on the foregoing solution, after the first communication apparatus obtains the performance monitoring information of the at least one AI model based on the first information, the first communication apparatus may perform noise perturbation on the model parameter of the first AI model, to obtain the updated model parameter of the first AI model. In this manner, computational power of a communication node can be applied to processing of an AI model, and the AI model can be updated based on performance monitoring information of the AI model.

[0009] In addition, a process in which the first communication apparatus updates a parameter of an AI model is triggered based on the performance monitoring information of the at least one AI model, and in the process of updating the parameter of the AI model, the first communication apparatus obtains an updated parameter of the AI model through noise perturbation. In this manner, the solution can be applied to a continual learning scenario. In addition, by performing noise perturbation on the parameter of the AI model, reduction or even disappear of plasticity in the AI model during long-term continual learning can be alleviated or avoided, thereby improving adaptability of continual learning and enhancing robustness of continual learning.

[0010] In this application, the AI model may be replaced by another term, for example, a neural network, a neural network model, an AI neural network model, a machine learning model, or an AI processing model.

[0011] In this application, noise perturbation may be replaced by another term, for example, noise addition, perturbation, noise perturbation, noise interference, or random perturbation.

[0012] In a possible implementation of the first aspect, the method further includes: The first communication apparatus sends second information based on the first information, where the second information is obtained by performing noise perturbation on a first parameter, the second information is used to update a second AI model, and the second AI model is deployed in a second communication apparatus.

[0013] Based on the foregoing solution, the first communication apparatus may further send,

based on the first information, the second information obtained by performing noise perturbation on the first parameter, so that a receiver of the second information (for example, the second communication apparatus) can obtain the second information obtained by performing noise perturbation on the first parameter, and update the second AI model based on the second information. In this manner, reduction or even disappear of plasticity in the second AI model during long-term continual learning can be alleviated or avoided, thereby improving adaptability of continual learning and enhancing robustness of continual learning.

[0014] In a possible implementation of the first aspect, the at least one AI model includes the first AI model and/or the second AI model, the first AI model is associated with the second AI model, and the second AI model is deployed in the second communication apparatus.

[0015] Based on the foregoing solution, the first information indicates the performance monitoring information of the at least one AI model, and when the performance monitoring information indicated by the first information meets a condition (for example, a first condition and/or a second condition described below), the first information may be used to trigger noise perturbation on the model parameter of the first AI model and/or a model parameter of the second AI model. Correspondingly, the at least one AI model may include the first AI model and/or the second AI model. In this manner, the first communication apparatus can trigger noise perturbation of an AI model based on performance of an associated AI model.

[0016] Optionally, that the first AI model is associated with the second AI model includes any one of the following:

the first AI model is a common model and the second AI model is a personalized model;

the first AI model is a personalized model and the second AI model is a common model;

an input of the first AI model includes an output of the second AI model;

an input of the second AI model includes an output of the first AI model;

the first AI model is a teacher model and the second AI model is a student model; or

the first AI model is a student model and the second AI model is a teacher model.

[0017] In a possible implementation of the first aspect, the second information is obtained by processing a quantization result of the first parameter, or the second information is obtained by processing a transmission parameter of the first parameter.

[0018] Based on the foregoing solution, the second information may be obtained by performing noise perturbation in a plurality of manners, to improve flexibility of implementing the solution.

[0019] Optionally, the first parameter includes one or more of the following:

an updated model parameter of the second AI model, gradient information used to update the second AI model, intermediate feature information used to update the second AI model,

intermediate gradient information used to update the second AI model, an inference result of the first AI model, distillation loss information corresponding to the second AI model, or task loss information of the second AI model.

[0020] In a possible implementation of the first aspect, that the first communication apparatus sends the second information based on the first information includes: The first communication apparatus sends the second information based on the first information when a first condition is met, where the first condition includes at least one of the following:

training accuracy of the at least one AI model is lower than a first threshold;

test accuracy of the at least one AI model is lower than a second threshold;

system performance of an AI model system in which the at least one AI model is located is lower than a third threshold;

system performance of a communication system in which a communication apparatus deployed with the at least one AI model is located is lower than a fourth threshold;

a neural network input-output distribution shift of the at least one AI model is greater than a fifth threshold;

a neural network weight distribution of the at least one AI model is greater than a sixth threshold; or

a neural network weight variation of the at least one AI model is lower than a seventh threshold.

[0021] Based on the foregoing solution, when the first condition is met, the first communication apparatus may determine that current running of the second AI model may be abnormal. Therefore, the first communication apparatus may send the second information, so that the second communication apparatus can obtain the second information obtained by performing noise perturbation on the first parameter, and update the second AI model based on the second information. In this manner, reduction or even disappear of plasticity in the second AI model during long-term continual learning can be alleviated or avoided, thereby improving adaptability of continual learning and enhancing robustness of continual learning.

[0022] In a possible implementation of the first aspect, the method further includes: The first communication apparatus receives indication information indicating a second parameter, where the second parameter is used to process the first parameter to obtain the second information.

[0023] Based on the foregoing solution, the first communication apparatus may further determine the second parameter based on the received indication information, and process the first parameter based on the second parameter to obtain the second information. In other words, the first communication apparatus may determine the second parameter based on an indication of another communication apparatus (for example, the receiver of the second information), so that

the receiver of the second information can obtain the corresponding second information based on the specified second parameter.

[0024] It should be understood that the second parameter may be used to perform noise perturbation on the first parameter to obtain the second information. Correspondingly, the second parameter may be understood as a noise perturbation parameter, a noise addition parameter, a perturbation parameter, or the like.

[0025] In a possible implementation of the first aspect, the method further includes: The first communication apparatus sends request information for requesting the second parameter.

[0026] Based on the foregoing solution, the first communication apparatus may determine the second parameter through an interaction process of the request information and the indication information.

[0027] Optionally, the second parameter is a preconfigured parameter.

[0028] In a possible implementation of the first aspect, the method further includes: The first communication apparatus receives indication information indicating a third parameter, where the third parameter is used to process the model parameter of the first AI model to obtain the updated model parameter of the first AI model.

[0029] Based on the foregoing solution, the first communication apparatus may further determine the third parameter based on the received indication information, and perform noise perturbation on the model parameter of the first AI model based on the third parameter. In other words, the first communication apparatus may determine the third parameter based on an indication of another communication apparatus (for example, the receiver of the second information), so that the first communication apparatus can implement noise perturbation based on the specified third parameter.

[0030] It should be understood that the third parameter may be used to process the model parameter of the first AI model to obtain the updated model parameter of the first AI model. Correspondingly, the third parameter may be understood as a noise perturbation parameter, a noise addition parameter, a perturbation parameter, or the like.

[0031] In a possible implementation of the first aspect, the method further includes: The first communication apparatus sends request information for requesting the third parameter.

[0032] Based on the foregoing solution, the first communication apparatus may determine the third parameter through an interaction process of the request information and the indication information.

[0033] Optionally, the third parameter is a preconfigured parameter.

[0034] In a possible implementation of the first aspect, that the first communication apparatus performs noise perturbation on the model parameter of the first AI model based on the first

information, to obtain the updated model parameter of the first AI model includes: When a second condition is met, the first communication apparatus performs noise perturbation on the model parameter of the first AI model based on the first information, to obtain the updated model parameter of the first AI model.

[0035] The second condition includes at least one of the following:

the training accuracy of the at least one AI model is lower than an eighth threshold;

the test accuracy of the at least one AI model is lower than a ninth threshold;

the system performance of the AI model system in which the at least one AI model is located is lower than a tenth threshold;

the system performance of the communication system in which the communication apparatus deployed with the at least one AI model is located is lower than an eleventh threshold;

the neural network input-output distribution shift of the at least one AI model is greater than a twelfth threshold;

the neural network weight distribution of the at least one AI model is greater than a thirteenth threshold; or

the neural network weight variation of the at least one AI model is lower than a fourteenth threshold.

[0036] Based on the foregoing solution, when the second condition is met, the first communication apparatus may determine that current running of the first AI model may be abnormal. Therefore, the first communication apparatus may process the model parameter of the first AI model to obtain the updated model parameter of the first AI model, to update the first AI model. In this manner, reduction or even disappear of plasticity in the first AI model during long-term continual learning can be alleviated or avoided, thereby improving adaptability of continual learning and enhancing robustness of continual learning.

[0037] In a possible implementation of the first aspect, that the first communication apparatus obtains the first information includes: The first communication apparatus receives the first information; or the first communication apparatus obtains a measurement parameter of the first AI model and/or a measurement parameter of the second AI model, and determines the first information based on the measurement parameter of the first AI model and/or the measurement parameter of the second AI model.

[0038] Based on the foregoing solution, the first communication apparatus may obtain the first information in the foregoing plurality of manners, to improve flexibility of implementing the solution.

[0039] A second aspect of this application provides a communication method. The method is applied to a first communication apparatus. The first communication apparatus may be a

communication device (for example, a terminal device or a network device), or the first communication apparatus may be some components (for example, a processor, a chip, or a system on chip) in the communication device, or the first communication apparatus may be a logic module or software that can implement all or some functions of the communication device. In the method, the first communication apparatus obtains first information, where the first information indicates performance monitoring information of at least one AI model, and the at least one AI model is associated with a second AI model deployed in a second communication apparatus; and the first communication apparatus sends second information based on the first information, where the second information is obtained by performing noise perturbation on a first parameter, and the second information is used to update the second AI model.

[0040] Based on the foregoing solution, after the first communication apparatus obtains the performance monitoring information of the at least one AI model based on the first information, the first communication apparatus may send the second information obtained by performing noise perturbation on the first parameter, so that a receiver of the second information (for example, the second communication apparatus) can obtain the second information obtained by performing noise perturbation on the first parameter, and update the second AI model based on the second information. In this manner, reduction or even disappear of plasticity in the second AI model during long-term continual learning can be alleviated or avoided, thereby improving adaptability of continual learning and enhancing robustness of continual learning.

[0041] In a possible implementation of the second aspect, the at least one AI model includes a first AI model and/or the second AI model, the first AI model is associated with the second AI model, and the second AI model is deployed in the second communication apparatus.

[0042] Based on the foregoing solution, the first information indicates the performance monitoring information of the at least one AI model, and when the first information is used to trigger the indicated performance monitoring information to meet a condition (for example, a first condition and/or a second condition described below), the first information may be used to perform noise perturbation on a model parameter of the first AI model and/or a model parameter of the second AI model. Correspondingly, the at least one AI model may include the first AI model and/or the second AI model. In this manner, the first communication apparatus can trigger noise perturbation of an AI model based on performance of an associated AI model.

[0043] Optionally, that the first AI model is associated with the second AI model includes any one of the following:

the first AI model is a common model and the second AI model is a personalized model;

the first AI model is a personalized model and the second AI model is a common model;

an input of the first AI model includes an output of the second AI model;

an input of the second AI model includes an output of the first AI model;

the first AI model is a teacher model and the second AI model is a student model; or

the first AI model is a student model and the second AI model is a teacher model.

[0044] In a possible implementation of the second aspect, the second information is obtained by processing a quantization result of the first parameter, or the second information is obtained by processing a transmission parameter of the first parameter.

[0045] Based on the foregoing solution, the second information may be obtained by performing noise perturbation in a plurality of manners, to improve flexibility of implementing the solution.

[0046] Optionally, the first parameter includes one or more of the following:

an updated model parameter of the second AI model, gradient information used to update the second AI model, intermediate feature information used to update the second AI model, intermediate gradient information used to update the second AI model, an inference result of the first AI model, distillation loss information corresponding to the second AI model, or task loss information of the second AI model.

[0047] In a possible implementation of the second aspect, that the first communication apparatus sends the second information based on the first information includes: The first communication apparatus sends the second information based on the first information when a first condition is met.

[0048] The first condition includes at least one of the following:

training accuracy of the at least one AI model is lower than a first threshold;

test accuracy of the at least one AI model is lower than a second threshold;

system performance of an AI model system in which the at least one AI model is located is lower than a third threshold;

system performance of a communication system in which a communication apparatus deployed with the at least one AI model is located is lower than a fourth threshold;

a neural network input-output distribution shift of the at least one AI model is greater than a fifth threshold;

a neural network weight distribution of the at least one AI model is greater than a sixth threshold; or

a neural network weight variation of the at least one AI model is lower than a seventh threshold.

[0049] Based on the foregoing solution, when the first condition is met, the first communication apparatus may determine that current running of the second AI model may be abnormal. Therefore, the first communication apparatus may send the second information, so that

the second communication apparatus can obtain the second information obtained by performing noise perturbation on the first parameter, and update the second AI model based on the second information. In this manner, reduction or even disappear of plasticity in the second AI model during long-term continual learning can be alleviated or avoided, thereby improving adaptability of continual learning and enhancing robustness of continual learning.

[0050] In a possible implementation of the second aspect, the method further includes: The first communication apparatus receives indication information indicating a second parameter, where the second parameter is used to process the first parameter to obtain the second information.

[0051] Based on the foregoing solution, the first communication apparatus may further determine the second parameter based on the received indication information, and process the first parameter based on the second parameter to obtain the second information. In other words, the first communication apparatus may determine the second parameter based on an indication of another communication apparatus (for example, the receiver of the second information), so that the receiver of the second information can obtain the corresponding second information based on the specified second parameter.

[0052] It should be understood that the second parameter may be used to perform noise perturbation on the first parameter to obtain the second information. Correspondingly, the second parameter may be understood as a noise perturbation parameter, a noise addition parameter, a perturbation parameter, or the like.

[0053] In a possible implementation of the second aspect, the method further includes: The first communication apparatus sends request information for requesting the second parameter.

[0054] Based on the foregoing solution, the first communication apparatus may determine the second parameter through an interaction process of the request information and the indication information.

[0055] Optionally, the second parameter is a preconfigured parameter.

[0056] In a possible implementation of the second aspect, that the first communication apparatus obtains the first information includes: The first communication apparatus receives the first information; or the first communication apparatus obtains a measurement parameter of the first AI model and/or a measurement parameter of the second AI model, and determines the first information based on the measurement parameter of the first AI model and/or the measurement parameter of the second AI model.

[0057] Based on the foregoing solution, the first communication apparatus may obtain the first information in the foregoing plurality of manners, to improve flexibility of implementing the solution.

[0058] A third aspect of this application provides a communication method. The method is

applied to a second communication apparatus. The second communication apparatus may be a communication device (for example, a terminal device or a network device), or the second communication apparatus may be some components (for example, a processor, a chip, or a system on chip) in the communication device, or the second communication apparatus may be a logic module or software that can implement all or some functions of the communication device. In the method, the second communication apparatus receives second information, where the second information is obtained by performing noise perturbation on a first parameter, and the second information is used to update a second AI model deployed in the second communication apparatus; and the second communication apparatus updates the second AI model based on the second information.

[0059] Based on the foregoing solution, the second information received by the second communication apparatus is obtained by performing noise perturbation on the first parameter, and the second communication apparatus can update the second AI model based on the second information. In other words, in a process in which the second communication apparatus updates the second AI model, the second information used to update the second AI model is obtained by performing noise perturbation on the first parameter. In this manner, reduction or even disappear of plasticity in the second AI model during long-term continual learning can be alleviated or avoided, thereby improving adaptability of continual learning and enhancing robustness of continual learning.

[0060] In a possible implementation of the third aspect, the second information is obtained by processing a quantization result of the first parameter, or the second information is obtained by processing a transmission parameter of the first parameter.

[0061] Based on the foregoing solution, the second information may be obtained by performing noise perturbation in a plurality of manners, to improve flexibility of implementing the solution.

[0062] Optionally, the first parameter includes one or more of the following:

an updated model parameter of the second AI model, gradient information used to update the second AI model, intermediate feature information used to update the second AI model, intermediate gradient information used to update the second AI model, an inference result of a first AI model, distillation loss information corresponding to the second AI model, or task loss information of the second AI model.

[0063] A fourth aspect of this application provides a communication apparatus. The communication apparatus is a first communication apparatus, and the communication apparatus includes a transceiver unit. The processing unit is configured to obtain first information, where the first information indicates performance monitoring information of at least one AI model, and the

at least one AI model is associated with a first AI model deployed in the first communication apparatus; and the processing unit is further configured to perform noise perturbation on a model parameter of the first AI model based on the first information, to obtain an updated model parameter of the first AI model.

[0064] In the fourth aspect of this application, a composition module of the communication apparatus may be further configured to: perform the steps performed in the possible implementations of the first aspect, and achieve corresponding technical effects. For details, refer to the first aspect. Details are not described herein again.

[0065] A fifth aspect of this application provides a communication apparatus. The communication apparatus is a first communication apparatus, and the communication apparatus includes a transceiver unit and a processing unit. The processing unit is configured to obtain first information, where the first information indicates performance monitoring information of at least one AI model, and the at least one AI model is associated with a second AI model deployed in a second communication apparatus. The transceiver unit is configured to send second information based on the first information, where the second information is obtained by performing noise perturbation on a first parameter, and the second information is used to update the second AI model.

[0066] In the fifth aspect of this application, a composition module of the communication apparatus may be further configured to: perform the steps performed in the possible implementation of the second aspect, and achieve corresponding technical effects. For details, refer to the second aspect. Details are not described herein again.

[0067] A sixth aspect of this application provides a communication apparatus. The communication apparatus is a second communication apparatus, and the communication apparatus includes a transceiver unit and a processing unit. The transceiver unit is configured to receive second information, where the second information is obtained by performing noise perturbation on a first parameter, and the second information is used to update a second AI model deployed in the second communication apparatus. The processing unit is configured to update the second AI model based on the second information.

[0068] In the sixth aspect of this application, a composition module of the communication apparatus may be further configured to: perform the steps performed in the possible implementation of the third aspect, and achieve corresponding technical effects. For details, refer to the third aspect. Details are not described herein again.

[0069] A seventh aspect of this application provides a communication apparatus, including at least one processor. The at least one processor is coupled to a memory. The memory is configured to store a program or instructions. The at least one processor is configured to execute the program or the instructions, so that the communication apparatus implements the method according to any

possible implementation of any one of the first aspect to the third aspect. Optionally, the communication apparatus may include the memory.

[0070] An eighth aspect of this application provides a communication apparatus, including at least one logic circuit and an input/output interface. The logic circuit is configured to perform the method according to any possible implementation of any one of the first aspect to the third aspect.

[0071] A ninth aspect of this application provides a communication system. The communication system includes the foregoing first communication apparatus and the foregoing second communication apparatus.

[0072] A tenth aspect of this application provides a computer-readable storage medium. The storage medium is configured to store one or more computer-executable instructions. When the computer-executable instructions are executed by a processor, the processor performs the method according to any possible implementation of any one of the first aspect to the third aspect.

[0073] An eleventh aspect of this application provides a computer program product (or referred to as a computer program). When the computer program in the computer program product is executed by a processor, the processor performs the method according to any possible implementation of any one of the first aspect to the third aspect.

[0074] A twelfth aspect of this application provides a system on chip. The system on chip includes at least one processor, configured to support a communication apparatus in implementing the method according to any possible implementation of any one of the first aspect to the third aspect.

[0075] In a possible design, the system on chip may further include a memory, and the memory is configured to store program instructions and data that are necessary for the communication apparatus. The system on chip may include a chip, or may include a chip and another discrete component. Optionally, the system on chip further includes an interface circuit, and the interface circuit provides program instructions and/or data for the at least one processor.

[0076] For technical effects brought by any design in the fourth aspect to the twelfth aspect, refer to the technical effects brought by different designs in the first aspect to the third aspect. Details are not described herein again.

BRIEF DESCRIPTION OF DRAWINGS

[0077] FIG. 1a to FIG. 1c are diagrams of communication systems according to this application;

[0078] FIG. 2a to FIG. 2g are diagrams of AI processing processes according to this application;

[0079] FIG. 3 is a diagram of interaction in a communication method according to this application;

[0080] FIG. 4a to FIG. 4c are diagrams of model processing according to this application;

[0081] FIG. 5 is a diagram of interaction in a communication method according to this application; and

[0082] FIG. 6 to FIG. 10 are diagrams of communication apparatuses according to this application.

DESCRIPTION OF EMBODIMENTS

[0083] First, some terms in embodiments of this application are explained and described, to facilitate understanding of a person skilled in the art.

[0084] (1) A terminal device may be a wireless terminal device that can receive scheduling and indication information of a network device. The wireless terminal device may be a device providing voice and/or data connectivity for a user, a handheld device having a wireless connection function, or another processing device connected to a wireless modem.

[0085] The terminal device may communicate with one or more core networks or the internet through a radio access network (radio access network, RAN). The terminal device may be a mobile terminal device like a mobile phone (or referred to as a "cellular" phone or a mobile phone (mobile phone)), a computer, and a data card. For example, the terminal device may be a portable, pocket-sized, handheld, computer built-in, or vehicle-mounted mobile apparatus that exchanges voices and/or data with the radio access network. For example, the terminal device may be a device like a personal communication service (personal communication service, PCS) phone, a cordless phone, a session initiation protocol (session initiation protocol, SIP) phone, a wireless local loop (wireless local loop, WLL) station, a personal digital assistant (personal digital assistant, PDA), a tablet computer (tablet or pad), or a computer having a wireless transceiver function. The wireless terminal device may also be referred to as a system, a subscriber unit (subscriber unit), a subscriber station (subscriber station), a mobile station/mobile console (mobile station, MS), a remote station (remote station), an access point (access point, AP), a remote terminal (remote terminal), an access terminal (access terminal), a user terminal (user terminal), a user agent (user agent), a subscriber station (subscriber station, SS), customer premises equipment (customer premises equipment, CPE), a terminal (terminal), user equipment (user equipment, UE), a mobile terminal (mobile terminal, MT), or the like.

[0086] By way of example, but not limitation, in embodiments of this application, the terminal device may alternatively be a wearable device. The wearable device may also be referred to as a

wearable intelligent device, an intelligent wearable device, or the like, and is a general term of wearable devices that are intelligently designed and developed for daily wear by using a wearable technology, for example, glasses, gloves, watches, clothes, and shoes. The wearable device is a portable device that can be directly worn on the body or integrated into clothes or an accessory of a user. The wearable device is not only a hardware device, but also implements a powerful function through software support, data exchange, and cloud interaction. In a broad sense, wearable intelligent devices include full-featured and large-sized devices that can implement all or some functions without depending on smartphones, for example, smart watches or smart glasses, and include devices that dedicated to only one type of application function and need to collaboratively work with other devices such as smartphones, for example, various smart bands, smart helmets, or smart jewelry for monitoring physical signs.

[0087] The terminal may alternatively be an uncrewed aerial vehicle, a robot, a terminal in device-to-device (device-to-device, D2D) communication, a terminal in vehicle-to-everything (vehicle-to-everything, V2X), a virtual reality (virtual reality, VR) terminal device, an augmented reality (augmented reality, AR) terminal device, a wireless terminal in industrial control (industrial control), a wireless terminal in self driving (self driving), a wireless terminal in telemedicine (telemedicine or telehealth services), a wireless terminal in smart grid (smart grid), a wireless terminal in transportation safety (transportation safety), a wireless terminal in smart city (smart city), a wireless terminal in smart home (smart home), or the like.

[0088] In addition, the terminal device may alternatively be a terminal device in a communication system (for example, a 5G Advanced or 6th generation (6th generation, 6G) communication system) evolved after a 5th generation (5th generation, 5G) communication system, a terminal device in a future evolved public land mobile network (public land mobile network, PLMN), or the like. For example, a 5G Advanced or 6G network may further extend a form and a function of a 5G communication terminal, and a 6G terminal includes but is not limited to a vehicle, a cellular network terminal (integrating a function of a satellite terminal), an uncrewed aerial vehicle, and an internet of things (internet of things, IoT) device.

[0089] In embodiments of this application, the terminal device may further obtain an artificial intelligence (artificial intelligence, AI) service provided by the network device. Optionally, the terminal device may further have an AI processing capability.

[0090] (2) A network device may be a device in a wireless network. For example, the network device may be a RAN node (or device) connecting a terminal device to the wireless network, and may also be referred to as a base station. Currently, some examples of the RAN device are a base station (base station), an evolved NodeB (evolved NodeB, eNodeB), a base station gNB (gNodeB) in a 5G communication system, a transmission reception point (transmission reception point, TRP),

an evolved NodeB (evolved NodeB, eNB), a radio network controller (radio network controller, RNC), a NodeB (NodeB, NB), a home base station (for example, a home evolved NodeB, or a home NodeB, HNB), a baseband unit (baseband unit, BBU), a wireless fidelity (wireless fidelity, Wi-Fi) access point (access point, AP), and the like. In addition, in a network structure, the network device may include a central unit (central unit, CU) node, a distributed unit (distributed unit, DU) node, or a RAN device including a CU node and a DU node.

[0091] Optionally, the RAN node may alternatively be a macro base station, a micro base station, an indoor base station, a relay node, or a donor node, or may be a radio controller in a cloud radio access network (cloud radio access network, CRAN) scenario. The RAN node may alternatively be a server, a wearable device, a vehicle, a vehicle-mounted device, or the like. For example, an access network device in a vehicle-to-everything (vehicle to everything, V2X) technology may be a road side unit (road side unit, RSU).

[0092] In another possible scenario, a plurality of RAN nodes coordinate to assist the terminal in implementing radio access, and different RAN nodes separately implement some functions of a base station. For example, the RAN node 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 baseband unit (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), a radio head (radio head, RH), 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 alternatively have different names, but a person skilled in the art may understand meanings of the names. For example, in an open access network (open RAN, O-RAN or ORAN) 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, or the RU in this application may be implemented by using a software module, a hardware module, or a combination of a software module and a hardware module.

[0094] Communication between the access network device and the terminal device complies with a specific protocol layer structure. Protocol layers may include a control plane protocol layer and a user plane protocol layer. The control plane protocol layer may include at least one of the following: a radio resource control (radio resource control, RRC) layer, a packet data convergence

protocol (packet data convergence protocol, PDCP) layer, a radio link control (radio link control, RLC) layer, a media access control (media access control, MAC) layer, a physical (physical, PHY) layer, or the like. The user plane protocol layer may include at least one of the following: a service data adaptation protocol (service data adaptation protocol, SDAP) layer, a PDCP layer, an RLC layer, a MAC layer, a physical layer, or the like.

[0095] For a correspondence between network elements in the ORAN system and protocol layer functions that may be implemented by the network elements, refer to Table 1.

Table 1

ORAN network element	3GPP protocol layer function
O-CU-CP	RRC+PCDP-control plane (PDCP-C)
O-CU-UP	SDAP+PCDP-user plane (PDCP-U)
O-DU	RLC+MAC+PHY-high
O-RU	PHY-low

[0096] The network device may be another apparatus that provides a wireless communication function for the terminal device. A specific technology and a specific device form that are used by the network device are not limited in embodiments of this application. For ease of description, this is not limited in embodiments of this application.

[0097] The network device may further include a core network device. For example, the core network device includes network elements such as a mobility management entity (mobility management entity, MME), a home subscriber server (home subscriber server, HSS), a serving gateway (serving gateway, S-GW), a policy and charging rules function (policy and charging rules function, PCRF), and a public data network gateway (public data network gateway, PDN gateway or P-GW) in a 4th generation (4th generation, 4G) network, and an access and mobility management function (access and mobility management function, AMF), a user plane function (user plane function, UPF), and a session management function (session management function, SMF) in a 5G network. In addition, the core network device may further include another core network device in the 5G network and a next-generation network of the 5G network.

[0098] In embodiments of this application, the network device may alternatively be a network node having an AI capability, and may provide an AI service for a terminal or another network device, for example, may be an AI node, a computational power node, a RAN node having an AI capability, or a core network element having an AI capability on a network side (an access network or a core network).

[0099] 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, for example, a system on chip, that can support the network device in implementing the function. The apparatus may be disposed in the network device. In the technical solutions provided in embodiments of this application, description is made by using an example in which the apparatus configured to implement the function of the network device is the network device.

[00100] (3) Configuration and preconfiguration: In this application, both the configuration and the preconfiguration are used. The configuration means that a network device/a server sends configuration information of some parameters or values of the parameters to a terminal through a message or signaling, so that the terminal determines, based on the values or the information, a communication parameter or a resource used during transmission. Similar to the configuration, the preconfiguration may be parameter information or a parameter value negotiated by a network device/a server with a terminal device in advance, or may be parameter information or a parameter value used by a base station/a network device or a terminal device as specified in a standard protocol, or may be parameter information or a parameter value pre-stored in a base station/a server or a terminal device. This is not limited in this application.

[00101] Further, these values and parameters may be changed or updated.

[00102] (4) Terms "system" and "network" in embodiments of this application may be used interchangeably. "A plurality of" means two or more than two. The term "and/or" describes an association relationship between associated objects, and indicates that three relationships may exist. For example, A and/or B may indicate the following cases: Only A exists, both A and B exist, and only B exists, where A and B may be singular or plural. The character "/" generally indicates an "or" relationship between the associated objects. "At least one of the following items (pieces)" or a similar expression thereof indicates any combination of these items, including a single item (piece) or any combination of a plurality of items (pieces). For example, "at least one of A, B, and C" includes A, B, C, AB, AC, BC, or ABC. In addition, unless otherwise specified, ordinal numbers such as "first" and "second" in embodiments of this application are used to distinguish between a plurality of objects, and are not used to limit a sequence, a time sequence, priorities, or importance of the plurality of objects.

[00103] (5) "Sending" and "receiving" in embodiments of this application indicate signal transmission directions. For example, "sending information to XX" may be understood as that a destination of the information is XX, and may include direct sending through an air interface, or include indirect sending by another unit or module through an air interface. "Receiving information from YY" may be understood as that a source of the information is YY, and may include direct receiving from YY through an air interface, or may include indirect receiving from YY via another unit or module through an air interface. "Sending" may also be understood as "outputting" of a

chip interface, and "receiving" may also be understood as "inputting" of the chip interface.

[00104] In other words, sending and receiving may be performed between devices, for example, between a network device and a terminal device; or may be performed inside a device, for example, sending or receiving between components, modules, chips, software modules, or hardware modules inside the device through a bus, a trace, or an interface.

[00105] It may be understood that, necessary processing, such as encoding and modulation, may be performed on information between a source and a destination of information sending, but the destination may understand valid information from the source. Similar descriptions in this application may be understood similarly. Details are not described again.

[00106] (6) In embodiments of this application, "indication" may include a direct indication and an indirect indication, or may include an explicit indication and an implicit indication. Information indicated by specific information (for example, indication information in the following descriptions) is referred to as to-be-indicated information. In this case, during specific implementation, the to-be-indicated information may be indicated in a plurality of manners, for example, but not limited to: The to-be-indicated information may be directly indicated, for example, the to-be-indicated information or an index of the to-be-indicated information is indicated. Alternatively, the to-be-indicated information may be indirectly indicated by indicating other information, where there is an association relationship between the other information and the to-be-indicated information. Alternatively, only a part of the to-be-indicated information may be indicated, and a remaining part of the to-be-indicated information is known or pre-agreed upon. For example, specific information may be indicated by an arrangement sequence of a plurality of pieces of information that is pre-agreed upon (for example, predefined in a protocol), to reduce indication overheads to a specific extent. A specific indication manner is not limited in this application. It may be understood that, for a sender of the indication information, the indication information may indicate the to-be-indicated information, and for a receiver of the indication information, the indication information may be used to determine the to-be-indicated information.

[00107] In this application, unless otherwise specified, mutual reference may be made between same or similar parts in embodiments. In embodiments of this application and methods/designs/implementations in embodiments, unless otherwise specified or logic conflicts occur, terms and/or descriptions between different embodiments and between the methods/designs/implementations in embodiments are consistent and may be mutually referenced, and different embodiments and technical features in the methods/designs/implementations in embodiments may be combined to form a new embodiment, method, or implementation based on an internal logic relationship thereof. The following implementations of this application do not constitute a limitation on the protection scope of this application.

[00108] This application may be applied to a long term evolution (long term evolution, LTE) system, a new radio (new radio, NR) system, or a communication system (for example, 6G) evolved after 5G. The communication system includes at least one network device and/or at least one terminal device.

[00109] FIG. 1a is a diagram of a communication system according to this application. FIG. 1a illustrates one network device and six terminal devices. The six terminal devices are a terminal device 1, a terminal device 2, a terminal device 3, a terminal device 4, a terminal device 5, a terminal device 6, and the like. In the example shown in FIG. 1a, an example in which the terminal device 1 is a smart teacup, the terminal device 2 is a smart air conditioner, the terminal device 3 is a smart fuel dispenser, the terminal device 4 is a vehicle, the terminal device 5 is a mobile phone, and the terminal device 6 is a printer is used for description.

[00110] As shown in FIG. 1a, an AI configuration information sending entity may be the network device. AI configuration information receiving entities may be the terminal device 1 to the terminal device 6. In this case, the network device and the terminal device 1 to the terminal device 6 form a communication system. In the communication system, the terminal device 1 to the terminal device 6 may send data to the network device, and the network device needs to receive the data sent by the terminal device 1 to the terminal device 6. In addition, the network device may send configuration information to the terminal device 1 to the terminal device 6.

[00111] For example, in FIG. 1a, the terminal device 4 to the terminal device 6 may also form a communication system. The terminal device 5 serves as a network device, namely, an AI configuration information sending entity. The terminal device 4 and the terminal device 6 serve as terminal devices, namely, AI configuration information receiving entities. For example, in an internet of vehicles system, the terminal device 5 separately sends AI configuration information to the terminal device 4 and the terminal device 6, and receives data sent by the terminal device 4 and the terminal device 6. Correspondingly, the terminal device 4 and the terminal device 6 receive the AI configuration information sent by the terminal device 5, and send the data to the terminal device 5.

[00112] The communication system shown in FIG. 1a is used as an example. In addition to a communication-related service, an AI-related service may be performed between different devices (including between network devices, between a network device and a terminal device, and/or between terminal devices).

[00113] As shown in FIG. 1b, an example in which the network device is a base station is used. A communication-related service and an AI-related service may be performed between the base station and one or more terminal devices, and a communication-related service and an AI-related service may also be performed between different terminal devices.

[00114] As shown in FIG. 1c, an example in which terminal devices include a television and a mobile phone is used. A communication-related service and an AI-related service may also be performed between the television and the mobile phone.

[00115] The technical solutions provided in this application may be applied to a wireless communication system (for example, the system shown in FIG. 1a, FIG. 1b, or FIG. 1c). For example, an AI network element may be introduced to the communication system provided in this application to implement a part or all of AI-related operations. The AI network element may also be referred to as an AI node, an AI device, an AI entity, an AI module, an AI model, an AI unit, or the like. The AI network element may be built in a network element in a communication system. For example, the AI network element may be an AI module built in the access network device, the core network device, a cloud server, or an operation, administration and maintenance (operation, administration and maintenance, OAM), to implement an AI-related function. The OAM may serve as a network manager of the core network device and/or a network manager of the access network device. Alternatively, the AI network element may be an independently disposed network element in the communication system. Optionally, the terminal or the chip built in the terminal may alternatively include the AI entity, configured to implement the AI-related function.

[00116] The following briefly describes AI that may be used in this application.

[00117] AI enables machines to have human intelligence, for example, enables machines to use computer software and hardware to simulate some intelligent human behaviors. To implement the artificial intelligence, a machine learning method may be used. In the machine learning method, a machine obtains a model through learning (or training) by using training data. The model represents mapping from an input to an output. The model obtained through learning may be used for inference (or prediction). To be specific, the model may be used to predict an output corresponding to a given input. The output may also be referred to as an inference result (or a prediction result).

[00118] Machine learning may include supervised learning, unsupervised learning, and reinforcement learning. The unsupervised learning may also be referred to as non-supervised learning.

[00119] In terms of supervised learning, based on collected sample values and sample labels, a mapping relationship between the sample values and the sample labels is learned by using a machine learning algorithm, and the learned mapping relationship is expressed by using an AI model. A process of training a machine learning model is a process of learning the mapping relationship. In the training process, a sample value is input into the model to obtain a predicted value of the model, and a model parameter is optimized by calculating an error between the predicted value of the model and a sample label (ideal value). After the mapping relationship is

learned, a new sample label may be predicted by using the learned mapping. The mapping relationship learned through the supervised learning may include linear mapping or non-linear mapping. Learning tasks may be classified into a classification task and a regression task based on a type of a label.

[00120] In terms of unsupervised learning, an algorithm autonomously explores the internal pattern of samples based on collected sample values. For a specific type of algorithm of the unsupervised learning, a sample is used as a supervised signal. In other words, a model learns a mapping relationship between samples, which is referred to as self-supervised learning. During training, a model parameter is optimized by calculating an error between a predicted value of the model and the sample. The self-supervised learning may be used for signal compression and decompression restoration. Common algorithms include an autoencoder, a generative adversarial network, and the like.

[00121] The reinforcement learning is different from the supervised learning, and is an algorithm that learns a policy of resolving problems by interacting with an environment. Different from the supervised learning and the unsupervised learning, the reinforcement learning does not have clear "correct" action label data. The algorithm needs to interact with the environment to obtain a reward signal fed back by the environment and adjust a decision action to obtain a larger reward signal value. For example, in downlink power control, a reinforcement learning model adjusts a downlink transmit power of each user based on a total system throughput fed back by a wireless network, to obtain a higher system throughput. An objective of the reinforcement learning is also to learn a mapping relationship between an environment status and a better (for example, an optimal) decision action. However, a label of "correct action" cannot be obtained in advance. Therefore, a network cannot be optimized by calculating an error between an action and the "correct action". Reinforcement learning training is implemented through iterative interaction with the environment.

[00122] A neural network (neural network, NN) is a specific model in a machine learning technology. According to a universal approximation theorem, the neural network can theoretically approximate any continuous function, so that the neural network has a capability of learning any mapping. In a conventional communication system, rich expertise is required to design a communication module. However, in a neural network-based deep learning communication system, an implicit pattern structure may be automatically discovered from a large quantity of datasets and a mapping relationship between data may be established, to obtain performance better than that of a conventional modeling method.

[00123] An idea of the neural network comes from a neuron structure of brain tissue. For example, each neuron performs a weighted summation operation on input values of the neuron,

and outputs an operation result through an activation function.

[00124] FIG. 2a is a diagram of a neuron structure. It is assumed that inputs of a neuron are $x = [x_0, x_1, \dots, x_n]$, and weights corresponding to the inputs are respectively $w = [w_0, w_1, \dots, w_n]$, where n is a positive integer, and w_i and x_i may be values of any possible type, for example, a decimal, an integer (for example, 0, a positive integer, or a negative integer), or a complex number. w_i is used as a weight of x_i, and is used to weight x_i. A bias for performing weighted summation on the input values based on the weights is, for example, b. There may be a plurality of forms of activation function. It is assumed that an activation function of a neuron is $y = f(z) = \max(0, z)$. In this case, an output of the neuron is $y = f\left(\sum_{i=0}^{i=n} w_i * x_i + b\right) = \max(0, \sum_{i=0}^{i=n} w_i * x_i + b)$. For another example, if an activation function of a neuron is $y = f(z) = z$, an output of the neuron is $y = f\left(\sum_{i=0}^{i=n} w_i * x_i + b\right) = \sum_{i=0}^{i=n} w_i * x_i + b$. Herein, b may be a value of any possible type, for example, a decimal, an integer (for example, 0, a positive integer, or a negative integer), or a complex number. Activation functions of different neurons in the neural network may be the same or different.

[00125] In addition, the neural network usually includes a plurality of layers, and each layer may include one or more neurons. A depth and/or a width of the neural network are/is increased, so that an expression capability of the neural network can be improved, and a more powerful information extraction and abstraction modeling capability can be provided for a complex system. The depth of the neural network may be a quantity of layers included in the neural network, and a quantity of neurons included at each layer may be referred to as a width of the layer. In an implementation, the neural network includes an input layer and an output layer. The input layer of the neural network performs neuron processing on received input information, and transfers a processing result to the output layer. The output layer obtains an output result of the neural network. In another implementation, the neural network includes an input layer, a hidden layer, and an output layer. The input layer of the neural network performs neuron processing on received input information, and transfers a processing result to an intermediate hidden layer. The hidden layer performs calculation on the received processing result to obtain a calculation result. The hidden layer transfers the calculation result to the output layer or a next adjacent hidden layer. Finally, the output layer obtains an output result of the neural network. One neural network may include one hidden layer, or include a plurality of hidden layers that are sequentially connected. This is not limited.

[00126] The neural network is, for example, a deep neural network (deep neural network, DNN). Based on a network construction manner, the DNN may include a feedforward neural network (feedforward neural network, FNN), a convolutional neural network (convolutional neural network, CNN), and a recurrent neural network (recurrent neural network, RNN).

[00127] FIG. 2b is a diagram of an FNN network. A feature of the FNN network is that neurons at adjacent layers are fully connected to each other. Due to this feature, the FNN usually needs a large amount of storage space, resulting in high computational complexity.

[00128] The CNN is a neural network dedicated to processing data of a grid-like structure. For example, both time series data (for example, timeline discrete sampling) and image data (for example, two-dimensional discrete sampling) may be considered as data of a grid-like structure. The CNN performs a convolution operation by extracting partial information through a window with a fixed size, instead of performing an operation based on all input information at one time. This greatly reduces a computation workload of model parameters. In addition, based on different types of information extracted through the window (for example, a person and an object in a same image are information of different types), a different convolution kernel operation may be used for each window, so that the CNN can better extract a feature of input data.

[00129] The RNN is a DNN network using feedback time series information. Inputs of the RNN include a new input value at a current moment and an output value of the RNN at a previous moment. The RNN is suitable for obtaining sequence features having a time correlation, and is especially suitable for applications such as speech recognition and channel encoding and decoding.

[00130] In the foregoing model training process of machine learning, a loss function (loss function) may be defined. The loss function describes a gap or a difference between an output value of the model and an ideal target value. The loss function may be expressed in a plurality of forms, and a specific form of the loss function is not limited. The model training process may be considered as the following process: A part or all of parameters of the model are adjusted, so that a value of the loss function is less than a threshold or meets a target requirement.

[00131] The model may also be referred to as an AI model, a rule, another name, or the like. The AI model may be considered as a specific method for implementing an AI function. The AI model represents a mapping relationship or a function between an input and an output of the model. The AI function may include one or more of the following: data collection, model training (or model learning), model information release, model inference (or referred to as model inference, inference, prediction, or the like), model monitoring or model verification, inference result release, or the like. The AI function may also be referred to as an AI (AI-related) operation or an AI-related function.

[00132] The following describes an example of an implementation process of a fully connected neural network with reference to the accompanying drawings.

[00133] 1. A fully connected neural network is also referred to as a multilayer perceptron (multilayer perceptron, MLP).

[00134] As shown in FIG. 2c, one MLP includes one input layer (left side), one output layer

(right side), and a plurality of hidden layers (middle). Each layer of the MLP includes several nodes, which are referred to as neurons. Neurons at two adjacent layers are connected to each other.

[00135] Optionally, in consideration of neurons at two adjacent layers, an output h of a neuron at a lower layer is a weighted sum of all neurons x at an upper layer connected to the neuron at the lower layer, and may be expressed, by using an activation function, as:

$$h = f(wx + b).$$

[00136] Herein, w is a weight matrix, b is a bias vector, and f is the activation function.

[00137] Further, optionally, the output of the neural network may be recursively expressed as:

$$y = f_m(w_m f_{m-1}(\dots) + b_m).$$

[00138] Herein, m is an index of a neural network layer, m is greater than or equal to 1, m is less than or equal to M, and M is a total quantity of layers of the neural network.

[00139] In other words, the neural network may be understood as a mapping relationship between an input dataset and an output dataset. The neural network is usually initialized randomly, and a process of obtaining the mapping relationship from random w and b based on existing data is referred to as training of the neural network.

[00140] Optionally, a specific training manner is evaluating an output result of the neural network by using a loss function.

[00141] As shown in FIG. 2d, an error may be backpropagated, and neural network parameters (including w and b) can be iteratively optimized by using a gradient descent method until a loss function reaches a minimum value, namely, "a better point (for example, the optimal point)" in FIG. 2d. It may be understood that a neural network parameter corresponding to "the better point (for example, the optimal point)" in FIG. 2d may be used as a neural network parameter in trained AI model information.

[00142] Further, optionally, a gradient descent process may be expressed as:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \frac{\partial L}{\partial \boldsymbol{\theta}}.$$

[00143] Herein, $\boldsymbol{\theta}$ is a to-be-optimized parameter (including **w** and **b**), L is a loss function, η is a learning rate for controlling a gradient descent step, ∂ represents differentiation, and $\frac{\partial L}{\partial \boldsymbol{\theta}}$ represents taking a derivative of $\boldsymbol{\theta}$ with respect to L.

[00144] Further, optionally, a chain rule for obtaining a partial derivative is used in a backpropagation process.

[00145] As shown in FIG. 2e, a gradient of a parameter at a given layer may be obtained by recursive calculation of a gradient of a parameter at an immediately subsequent layer, and may be expressed as:

$$\frac{\partial L}{\partial w_{ij}} = \frac{\partial L}{\partial s_i} \frac{\partial s_i}{\partial w_{ij}}.$$

[00146] Herein, w_{ij} is a weight of a connection between a node j and a node i, and s_i is a weighted sum of inputs on the node i.

[00147] 2. Federated learning (federated learning, FL)

[00148] A concept of federated learning is proposed to effectively resolve difficulties faced by the current development of artificial intelligence. While ensuring user data privacy and security, the federated learning facilitates various edge devices and a server at a central end to collaborate to efficiently complete a learning task of a model.

[00149] As shown in FIG. 2f, an FL architecture is a most widely used training architecture in the FL field currently. A FedAvg algorithm is a basic algorithm of FL. An algorithm procedure thereof is roughly as follows:

[00150] (1) A central end initializes a to-be-trained model w_g^0 and broadcasts and sends the model to all client devices.

[00151] (2) In a $t \in [1, T]^{\text{th}}$ round, a client $k \in [1, K]$ trains a received global model w_g^{t-1} based on a local dataset $\mathcal{D}_k$ for E epochs to obtain a local training result w_k^t, and then reports the local training result to a central node.

[00152] (3) The central node aggregates local training results from all (or some) clients. It is assumed that a set of clients that upload local models in a tth round is $\mathcal{S}^t$. The central end performs weighted averaging by using a quantity of samples of a corresponding client as a weight to obtain a new global model. A specific update rule is $w_g^t = \sum_{k \in \mathcal{S}^t} \frac{D_k w_k^t}{\sum_{k \in \mathcal{S}^t} D_k}$. Then, the central end broadcasts and sends a global model w_g^t of a latest version to all client devices for a new round of training.

[00153] (4) Steps (2) and (3) are repeated until the model converges finally or a quantity of training rounds reaches an upper limit.

[00154] In addition to reporting the local model w_k^t, a trained local gradient g_k^t may also be reported. The central node averages local gradients, and updates the global model based on a direction of an averaged gradient.

[00155] It can be learned that, in the FL framework, a dataset exists on a distributed node, that is, the distributed node collects a local dataset, performs local training, and reports a local result (a model or a gradient) obtained through training to the central node. The central node does not have a dataset, and is only configured to: fuse training results of distributed nodes to obtain a global model, and deliver the global model to the distributed nodes.

[00156] 3. Decentralized learning: Different from the federated learning, another distributed learning architecture is decentralized learning.

[00157] As shown in FIG. 2g, a fully distributed system without a central node is considered. A design objective $f(x)$ of a decentralized learning system is usually an average value of

objectives $f_i(x)$ of all nodes, that is, $f(x) = \frac{1}{p}\sum_{i=1}^{p} f_i(x)$. Herein, p is a quantity of distributed nodes, and x is a to-be-optimized parameter. In machine learning, x is a parameter of a machine learning (for example, neural network) model. Each node calculates a local gradient $\nabla f_i(x)$ by using local data and a local objective $f_i(x)$, and then sends the local gradient to a communication-reachable neighboring node. After receiving gradient information sent by a neighboring node, any node may update a parameter x of the local model according to the following formula:

$$x_i^{k+1} = x_i^k - \alpha_k \left(\frac{1}{|P_i|}\sum_{j\in P_i} \nabla f_j\left(x_j^k\right)\right).$$

[00158] Herein, x_i^{k+1} represents a parameter of a local model after a $(k+1)^{\text{th}}$ update (k is a natural number) on an ith node, x_i^k represents a parameter of the local model after a kth update on the ith node (if k is 0, x_i^k represents a parameter of the local model that is not updated on the ith node), α_k represents an optimization coefficient, P_i is a set of neighboring nodes of the node i, and $|P_i|$ represents a quantity of elements in the set of neighboring nodes of the node i, namely, a quantity of neighboring nodes of the node i. Through information exchange between nodes, the decentralized learning system finally learns a unified model.

[00159] The technical solutions provided in this application may be applied to a wireless communication system (for example, the system shown in FIG. 1a, FIG. 1b, or FIG. 1c). In the wireless communication system, a communication node usually has signal receiving and sending capabilities and a computational capability. A network device having a computational capability is used as an example. The computational capability of the network device is mainly used to provide computational power support for signal receiving and sending capabilities (for example, sending and receiving a signal), to implement a communication task between the network device and another communication node.

[00160] However, in a communication network, in addition to providing computational power support for the foregoing communication task based on the computational capability of the communication node, the communication node may further have spare computational capabilities. Therefore, how to use these computational capabilities is a technical problem.

[00161] In a possible implementation, the communication node may serve as a participating node of an AI learning system, and the computational power of the communication node is applied to a specific phase of the AI learning system (for example, the AI learning system shown in FIG. 2f or FIG. 2g). Generally, in an AI learning system, an AI node may establish a dataset based on a task, complete model training offline, and perform inference after online deployment. However, after long-term continual learning, plasticity of the AI model may disappear, in other words, the AI model cannot learn knowledge of a new training sample. Consequently, performance of the AI model deteriorates, and robustness of continual learning declines. In a possible optimization

manner, some parameters may be selectively re-initialized during training, to ensure that the neural network can maintain a capability of learning a new sample. However, in this optimization manner, a complex function needs to be designed and a life cycle of each parameter needs to be counted to select a reset parameter set, which is complex. In addition, re-initialization of a parameter greatly affects instantaneous performance of the neural network. Consequently, the foregoing optimization manner cannot effectively resolve the foregoing problem.

[00162] To resolve the foregoing problem, this application provides a communication method and a related device. The following provides detailed descriptions with reference to the accompanying drawings.

[00163] FIG. 3 is a diagram of an implementation of a communication method according to this application. The method includes the following steps.

[00164] It should be noted that in FIG. 3 and FIG. 5 below, an example in which a first communication apparatus and another communication apparatus (for example, a second communication apparatus) are used as execution bodies of the interaction example is used to illustrate the method. However, the execution bodies of the interaction example are not limited in this application. For example, the communication apparatus may be a communication device (for example, a terminal device or a network device), or a chip, a baseband (baseband) chip, a modem (modem) chip, a system-on-chip (system-on-chip, SoC) chip including a modem core, a system-in-package (system-in-package, SIP) chip, a communication module, a system on chip, a processor, a logic module, or software in a communication device.

[00165] S301: The second communication apparatus sends first information, and correspondingly, the first communication apparatus receives the first information. The first information indicates performance monitoring information of at least one AI model, and the at least one AI model is associated with a first AI model deployed in the first communication apparatus.

[00166] S302: The first communication apparatus performs noise perturbation on a model parameter of the first AI model based on the first information, to obtain an updated model parameter of the first AI model.

[00167] In this application, the AI model may be replaced by another term, for example, a neural network, a neural network model, an AI neural network model, a machine learning model, or an AI processing model.

[00168] In this application, noise perturbation may be replaced by another term, for example, noise addition, perturbation, noise perturbation, noise interference, or random perturbation.

[00169] In a possible implementation, the first information obtained by the first communication apparatus indicates the performance monitoring information of the at least one AI model, the at

least one AI model includes the first AI model and/or a second AI model, the first AI model is associated with the second AI model, and the second AI model is deployed in the second communication apparatus. The first information may indicate the performance monitoring information of the at least one AI model, and in S302, the first communication apparatus may perform noise perturbation on the model parameter of the first AI model based on the first information, to obtain the updated model parameter of the first AI model. In other words, when the performance monitoring information indicated by the first information meets a condition (for example, a first condition and/or a second condition described below), the first information may be used to trigger noise perturbation on the model parameter of the first AI model and/or a model parameter of the second AI model. Correspondingly, the at least one AI model may include the first AI model and/or the second AI model. In this manner, the first communication apparatus can trigger noise perturbation of an AI model based on performance of an associated AI model.

[00170] It should be noted that S301 is optional. For the first communication apparatus, in addition to obtaining the first information by interacting with the second communication apparatus, the first communication apparatus may further obtain the first information in another manner. For example, the first communication apparatus may locally obtain the first information. For example, the first communication apparatus may obtain a measurement parameter of the first AI model and/or a measurement parameter of the second AI model, and determine the first information based on the measurement parameter of the first AI model and/or the measurement parameter of the second AI model.

[00171] In an example, a measurement parameter of an AI model (for example, the first AI model and/or the second AI model) may include an output of the AI model. In other words, the first communication apparatus may determine performance monitoring information of the AI model based on the output of the AI model. For example, the performance monitoring information may be determined based on a mathematical relationship (for example, at least one of a difference, a variance, or a standard deviation) between the output of the AI model and a preconfigured label.

[00172] In another example, when an output of an AI model (for example, the first AI model and/or the second AI model) is a communication parameter (for example, at least one of a bit rate for encoding and/or decoding, a sequence used for scrambling and/or descrambling, a modulation and/or demodulation order, a channel parameter, a communication rate, or a bit error rate), a measurement parameter of the AI model may include communication performance information (for example, one or more of a bit error rate, a block error rate, reference signal received quality (reference signal received power, RSRP), or a signal to interference plus noise ratio (signal to interference plus noise ratio, SINR)) for communication based on the communication parameter. In other words, the first communication apparatus may determine performance monitoring

information of the AI model based on the communication performance information and preconfigured performance information. For example, the performance monitoring information may be determined based on a mathematical relationship (for example, at least one of a difference, a variance, or a standard deviation) between the communication performance information and the preconfigured performance information.

[00173] Correspondingly, the first communication apparatus may obtain the measurement parameter of the AI model in a plurality of manners. For example, the first communication apparatus obtains the measurement parameter of the first AI model based on an output of the local first AI model. For another example, the first communication apparatus obtains the measurement parameter of the first AI model based on a process of communicating with another communication apparatus. For another example, the first communication apparatus obtains the measurement parameter of the second AI model based on a process of communicating with the second communication apparatus.

[00174] In a possible implementation, that the first AI model is associated with the second AI model includes any one of Manner A to Manner F below.

[00175] Manner A: The first AI model is a common model and the second AI model is a personalized model.

[00176] Manner B: The first AI model is a personalized model and the second AI model is a common model.

[00177] Manner C: An input of the first AI model includes an output of the second AI model.

[00178] Manner D: An input of the second AI model includes an output of the first AI model.

[00179] Manner E: The first AI model is a teacher model and the second AI model is a student model.

[00180] Manner F: The first AI model is a student model and the second AI model is a teacher model.

[00181] Manner A and Manner B may be understood as the federated learning scenario shown in FIG. 2f. In federated learning, there are a central node and distributed nodes. The central node stores a common model, and periodically aggregates personalized models uploaded by the distributed nodes. The distributed node periodically downloads the common model, and fine-tunes the common model by using local data to obtain the personalized model. In Manner A, the first communication apparatus may be a central node, and the second communication apparatus may be a distributed node. In Manner B, the second communication apparatus may be a central node, and the first communication apparatus may be a distributed node.

[00182] In an example, FIG. 4a is a diagram of an implementation of federated learning. An example in which a network device is a central node and terminal devices are distributed nodes is

used. The network device may store a common model, and a terminal device 1 to a terminal device 3 may store respective personalized models. The network device and the terminal devices may interact with each other through a model download process and a model upload process. Some or all AI models in the terminal device may participate in federated learning. In the example shown in FIG. 4a, both the AI model of the terminal device 1 and the AI model of the terminal device 2 may participate in federated learning, and the AI model of the terminal device 3 may include a part filled with a black pattern that participates in federated learning and a blank part that does not participate in federated learning.

[00183] Each device shown in FIG. 4a may serve as a first communication apparatus to perform noise perturbation on a model parameter of a locally deployed AI model. In other words, noise perturbation is performed on a model parameter of the common model or the personalized model, to enhance robustness of continuous federated learning to a distribution change.

[00184] In an example of noise perturbation, for any AI model, a standard deviation σ_i of noise perturbation and a model parameter set $\mathbb{P}$ ($\mathbb{P}$ represents a set, and p_i represents an element in the set) may be defined. $p_i = 1$ represents noise perturbation is performed on the parameter (where a value range of i is 0 to N–1, and N is a quantity of neural network parameters participating in noise perturbation), with a value of $n_i \sim \mathcal{N}(0, \sigma_i^2)$ (that is, the noise n_i follows a normal distribution $(0, \sigma_i^2)$, an average value is 0, and a variance is σ_i^2). $p_i = 0$ represents that perturbation does not need to be performed.

[00185] Optionally, the parameter on which noise perturbation is performed may be some or all model parameters of the AI model. For example, the some or all model parameters are determined based on the foregoing set $\mathbb{p}$. For example, the AI model has a total of 1000 (that is, N=1000) parameters, the set $\mathbb{P}$ is {1, 0, 0, 1...}, a sequence of 1000 elements in the set $\mathbb{P}$ corresponds to the 1000 parameters, and noise perturbation is performed on a model parameter corresponding to a position of p_i whose value is 1 in the set $\mathbb{P}$.

[00186] In another example of noise perturbation, for any AI model whose model parameter set is a set $\mathbb{p}$, noise perturbation satisfies the following:

$$e_i = e_i + n_i.$$

[00187] Herein, e_i represents an ith parameter of the model. When $p_i = 1$, $n_i \sim \mathcal{N}(0, \sigma_i^2)$ ∋ (that is, the noise n_i follows a normal distribution $(0, \sigma_i^2)$, an average value is 0, and a variance is σ_i^2). When $p_i = 0$, $n_i = 0$.

[00188] Optionally, in addition to following the normal distribution, the noise n_i may be implemented in another manner. For example, n_i follows a uniform distribution $n_i \sim \mathcal{U}(a, b)$, where a and b represent a lower bound and an upper bound.

[00189] Manner C and Manner D may be understood as a split learning scenario. In split

learning, an AI model may be split into two or more parts.

[00190] In an example, FIG. 4b is a diagram of split learning. In this example, an AI model is split into two parts. The two parts may be an encoder and a decoder, and the encoder and the decoder communicate with each other about an intermediate feature and an intermediate gradient. During training, data enters an encoding neural network, and after an intermediate feature is obtained, the intermediate feature is transmitted to a decoding neural network. The decoding neural network calculates a loss based on a label value, and backward updates the decoding neural network. In addition, the decoding neural network transmits an intermediate gradient to the encoding neural network, and the encoding neural network updates the encoding neural network based on the received intermediate gradient.

[00191] The encoder or the decoder shown in FIG. 4b may serve as the first AI model in the first communication apparatus to perform noise perturbation. In other words, noise may be added to some or all model parameters of the encoder or decoder, to enhance robustness of continuous split learning to a distribution change. Similarly, for a manner of performing noise perturbation on the some or all model parameters of the encoder or the decoder, refer to the descriptions of the foregoing example.

[00192] Manner E and Manner F may be understood as a knowledge distillation scenario. Knowledge distillation may enhance a loss of a smaller or simpler model (for example, a student model or a knowledge learning model) by using an output of a larger or more complex model (for example, a teacher model or a knowledge transfer model), to obtain better convergence performance or a faster convergence speed. Knowledge distillation may be considered as a model compression technology.

[00193] In an example, FIG. 4c is a diagram of knowledge distillation. In this example, models involved in a knowledge distillation process include a teacher model and a student model. Generally, the teacher model does not participate in training, but this is not limited herein. To be specific, a specific teacher model and a specific student model are not distinguished. The teacher model may alternatively be updated under guidance of the student model, or the two models are updated simultaneously, or there are a plurality of teacher models or a plurality of student models. The following uses the most common single-teacher-guided single-student model for learning as an example. In FIG. 4c, the teacher model and the student model separately perform inference to obtain inference results. A knowledge distillation loss (knowledge distillation loss, KD loss) may be obtained based on the inference results of the teacher model and the student model. A device in which the student model is deployed may obtain a task loss (task loss) based on a label, and then the device may update a model parameter of the student model based on the distillation loss and the task loss.

[00194] The teacher model or the student model shown in FIG. 4c may serve as the first AI model in the first communication apparatus to perform noise perturbation. In other words, noise may be added to some or all model parameters of the teacher model or the student model, to enhance robustness of a continuous knowledge distillation process to a distribution change. Similarly, for a manner of performing noise perturbation on the some or all model parameters of the teacher model or the student model, refer to the descriptions of the foregoing example.

[00195] In a possible implementation, in S302, when a second condition is met, the first communication apparatus performs noise perturbation on the model parameter of the first AI model based on the first information, to obtain the updated model parameter of the first AI model, where the second condition includes at least one of the following:

training accuracy of the at least one AI model is lower than a first threshold;

test accuracy of the at least one AI model is lower than a second threshold;

system performance of an AI model system in which the at least one AI model is located is lower than a third threshold;

system performance of a communication system in which a communication apparatus deployed with the at least one AI model is located is lower than a fourth threshold;

a neural network input-output distribution shift of the at least one AI model is greater than a fifth threshold;

a neural network weight distribution of the at least one AI model is greater than a sixth threshold; or

a neural network weight variation of the at least one AI model is lower than a seventh threshold.

[00196] Optionally, the first communication apparatus may determine values of the first threshold to the seventh threshold in a plurality of manners. For example, one or more of the first threshold to the seventh threshold may be preconfigured. For another example, one or more of the first threshold to the seventh threshold may be configured by another device (for example, the second communication apparatus, a network device, or a server) for the first communication apparatus.

[00197] When the second condition is met, the first communication apparatus may determine that current running of the first AI model may be abnormal. Therefore, the first communication apparatus may process the model parameter of the first AI model to obtain the updated model parameter of the first AI model, to update the first AI model. In this manner, reduction or even disappear of plasticity in the first AI model during long-term continual learning can be alleviated or avoided, thereby improving adaptability of continual learning and enhancing robustness of continual learning.

[00198] Based on the solution shown in FIG. 3, after the first communication apparatus obtains the performance monitoring information of the at least one AI model based on the first information, in S302, the first communication apparatus may perform noise perturbation on the model parameter of the first AI model, to obtain the updated model parameter of the first AI model. In this manner, computational power of a communication node can be applied to processing of an AI model, and the AI model can be updated based on performance monitoring information of the AI model.

[00199] In addition, in S302, a process in which the first communication apparatus updates a parameter of an AI model is triggered based on the performance monitoring information of the at least one AI model, and in the process of updating the parameter of the AI model, the first communication apparatus obtains an updated parameter of the AI model through noise perturbation. In this manner, the solution can be applied to a continual learning scenario. In addition, by performing noise perturbation on the parameter of the AI model, reduction or even disappear of plasticity in the AI model during long-term continual learning can be alleviated or avoided, thereby improving adaptability of continual learning and enhancing robustness of continual learning.

[00200] In a possible implementation of the method shown in FIG. 3, before S302, the method further includes: The first communication apparatus receives indication information indicating a third parameter, where the third parameter is used to process the model parameter of the first AI model to obtain the updated model parameter of the first AI model. Specifically, the first communication apparatus may further determine the third parameter based on the received indication information, and perform noise perturbation on the model parameter of the first AI model based on the third parameter. In other words, the first communication apparatus may determine the third parameter based on an indication of another communication apparatus (for example, a receiver of second information), so that the first communication apparatus can implement noise perturbation based on the specified third parameter.

[00201] For example, the second communication apparatus may send, to the first communication apparatus based on the first information, the indication information indicating the third parameter. In other words, the second communication apparatus may send, when determining that the second condition is met, the indication information indicating the third parameter, so that the first communication apparatus obtains the third parameter and performs S302.

[00202] Optionally, the indication information indicating the third parameter may include at least one of the third parameter or an index of the third parameter (for an implementation of the index, refer to the manner shown in Table 2 or Table 3).

[00203] It should be understood that the third parameter may be used to process the model parameter of the first AI model to obtain the updated model parameter of the first AI model.

Correspondingly, the third parameter may be understood as a noise perturbation parameter, a noise addition parameter, a perturbation parameter, or the like. For example, the third parameter may be one or more of the sets $\mathbb{P}$, σ, and $\sigma_{\mathbb{i}}^{\mathbb{2}}$ described above.

[00204] Optionally, before the first communication apparatus receives the indication information indicating the third parameter, the method further includes: The first communication apparatus sends request information for requesting the third parameter. Therefore, the first communication apparatus may determine the third parameter through an interaction process of the request information and the indication information. For example, the first communication apparatus may send, based on the first information, the request information for requesting the third parameter. In other words, the first communication apparatus may send, when determining that the second condition is met, the request information for requesting the third parameter, to obtain the third parameter and perform S302.

[00205] Optionally, the third parameter is a preconfigured parameter.

[00206] In an example, the third parameter includes the set $\mathbb{P}$, and the set $\mathbb{P}$ may be determined based on Table 2.

Table 2

Index	Value of a set $\mathbb{p}$
0	Randomly select α% of all parameters, where α is a hyperparameter
1	Randomly select α% of parameters at an ith layer
2	Sequentially select α% of all parameters starting from the first layer
3	Reversely select α% of all parameters starting from the last layer
...	...

[00207] In an example, the third parameter includes σ, and σ may be determined based on Table 3.

Table 3

Index	σ
0	3.2
1	1
2	0.32
...	...

[00208] It can be learned from FIG. 3 and related descriptions that the first communication apparatus may perform, based on the performance monitoring information of the at least one AI model, noise perturbation on the model parameter of the first AI model deployed in the first communication apparatus, to enhance robustness of continual learning. The at least one AI model may be associated with the first AI model, or may be associated with the second AI model deployed in the second communication apparatus. In other words, the first communication apparatus may alternatively perform, based on the performance monitoring information of the at least one AI model, noise perturbation on the model parameter of the second AI model deployed in the second communication apparatus, to enhance robustness of continual learning. The following provides descriptions by using a process shown in FIG. 5.

[00209] FIG. 5 is another diagram of a communication method according to this application.

[00210] S501: A second communication apparatus sends first information, and correspondingly, a first communication apparatus receives the first information. The first information indicates performance monitoring information of at least one AI model, and the at least one AI model is associated with a first AI model deployed in the first communication apparatus.

[00211] It should be noted that S501 is optional. For the first communication apparatus, in addition to obtaining the first information by interacting with the second communication apparatus, the first communication apparatus may further obtain the first information in another manner. For a specific implementation process, refer to the foregoing S301 and related descriptions.

[00212] S502: The first communication apparatus sends second information based on the first information, and correspondingly, the second communication apparatus receives the second information. The second information is obtained by performing noise perturbation on a first parameter, the second information is used to update a second AI model, and the second AI model is deployed in the second communication apparatus.

[00213] Optionally, the first parameter includes one or more of the following: an updated model parameter of the second AI model, gradient information used to update the second AI model, intermediate feature information used to update the second AI model, intermediate gradient information used to update the second AI model, an inference result of the first AI model, distillation loss information corresponding to the second AI model, or task loss information of the second AI model.

[00214] In a possible implementation, the second information sent by the first communication apparatus in S502 may be implemented in a plurality of manners, and details are described below.

[00215] Manner 1: The second information is obtained by processing a quantization result of the first parameter.

[00216] Optionally, in Manner 1, when the second information is obtained by processing the

quantization result of the first parameter, after the second communication apparatus receives the second information in S502, the second communication apparatus may perform parsing processing (for example, the parsing processing may include one or more of channel equalization, demodulation, decoding, digital domain processing, or the like) on the received second information, and update the second AI model based on a parsing processing result.

[00217] Manner 2: The second information is obtained by processing a transmission parameter of the first parameter.

[00218] Optionally, in Manner 2, when the second information is obtained by processing the transmission parameter of the first parameter, the implementation process may be understood as an over-the-air computation process. To be specific, after the second communication apparatus receives the second information in S502, the second communication apparatus may not perform digital domain processing on the received second information, but directly update the second AI model based on original information of the received second information. For example, the first communication apparatus converts one or more parameter values included in the first parameter into one or more first symbols, performs noise perturbation on the one or more first symbols to obtain one or more second symbols, and then maps the one or more second symbols to an air interface resource (for example, at least one of a time domain resource, a frequency domain resource, or a spatial domain resource), to generate the second information. In this example, the first communication apparatus does not perform source coding on the first parameter at an application layer and then send the first parameter to a physical layer in a bit stream, where the physical layer performs channel encoding and symbol modulation and then generates a signal for sending. Instead, the first communication apparatus converts the first parameter into one or more symbols for transmission. Therefore, in a manner of transmitting the first parameter over a wireless air interface, a transmission speed of the first parameter can be increased, to improve communication efficiency, and increase a processing speed of performing a model update by the second communication apparatus based on the second information.

[00219] Optionally, the transmission parameter of the first parameter may include one or more of a transmit power parameter, a precoding parameter, or indication information indicating whether to perform processing in Manner 2 (the indication information may be used to control whether one or more communication apparatuses perform information transmission in Manner 2, so as to control, by controlling a quantity of accessed communication apparatuses, a signal-to-noise ratio of a signal obtained through over-the-air aggregation).

[00220] The following uses Manner 2 as an example to describe a process of sending the second information in different implementations of the AI model.

[00221] In an example, when the first AI model and the second AI model are implemented in

Manner A or Manner B, the first parameter may include one or more of the updated model parameter of the second AI model or the gradient information used to update the second AI model.

[00222] The process shown in FIG. 4a is used as an example. The first communication apparatus may be a network device. The first communication apparatus may perform noise perturbation on a model parameter or an update gradient of a common model (that is, the first parameter includes the model parameter or the update gradient of the common model) in a process of delivering the common model to obtain the second information, and send the second information in S502.

[00223] The process shown in FIG. 4a is still used as an example. The first communication apparatus may be any terminal device. The first communication apparatus may perform noise perturbation on a model parameter or an update gradient of a personalized model (that is, the first parameter includes the model parameter or the update gradient of the personalized model) in a process of uploading the personalized model to obtain the second information, and send the second information in S502.

[00224] In other words, in a model download or model upload process, lossless transmission may be changed to lossy transmission. That is, transmission noise is superimposed on the model parameter or the update gradient during upload or download, to achieve an effect of noise perturbation.

[00225] In another example, when the first AI model and the second AI model are implemented in Manner C or Manner D, the first parameter may include one or more of the intermediate feature information used to update the second AI model or the intermediate gradient information used to update the second AI model.

[00226] Similarly, the process shown in FIG. 4b is used as an example. The first communication apparatus may perform noise perturbation when the intermediate feature and the intermediate gradient are transmitted, so that a transmission manner of the intermediate feature and the intermediate gradient of the encoder or the decoder changes from lossless transmission to lossy transmission. In other words, transmission noise is superimposed on the intermediate feature and the intermediate gradient during upload/download, to achieve an effect of noise perturbation.

[00227] In another example, when the first AI model and the second AI model are implemented in Manner E or Manner F, the first parameter may include one or more of the inference result of the first AI model, the distillation loss information corresponding to the second AI model, or the task loss information of the second AI model.

[00228] Similarly, the process shown in FIG. 4c is used as an example. It is assumed that a calculation formula of the distillation loss is $\mathrm{loss}_{\mathrm{KD}} = \mathrm{L}(\mathrm{f_t}, \mathrm{f_s})$, and $\mathrm{f_t}$ and $\mathrm{f_s}$ are respectively an input of the teacher model and an input of the student model during calculation of the distillation

loss, for example, an intermediate feature of the AI model or an output result of the AI model. In a knowledge distillation training process, calculation of the distillation loss may be performed on a same node as the teacher model, or may be performed on a same node as the student model, or may be separately performed on another node. Therefore, for knowledge distillation, lossy transmission may be used to transmit f_t and/or f_s, to achieve an effect of noise perturbation.

[00229] In a possible implementation, in S502, the first communication apparatus sends the second information based on the first information when a first condition is met, where the first condition includes at least one of the following:

training accuracy of the at least one AI model is lower than an eighth threshold;

test accuracy of the at least one AI model is lower than a ninth threshold;

system performance of an AI model system in which the at least one AI model is located is lower than a tenth threshold;

system performance of a communication system in which a communication apparatus deployed with the at least one AI model is located is lower than an eleventh threshold;

a neural network input-output distribution shift of the at least one AI model is greater than a twelfth threshold;

a neural network weight distribution of the at least one AI model is greater than a thirteenth threshold; or

a neural network weight variation of the at least one AI model is lower than a fourteenth threshold.

[00230] Optionally, the first condition may be the same as a second condition. For example, an ith (a value of i ranges from 1 to 7) one of a first threshold to a seventh threshold is the same as an ith one of the eighth threshold to the fourteenth threshold.

[00231] Optionally, the first condition may be different from the second condition. For example, the ith (the value of i ranges from 1 to 7) one of the first threshold to the seventh threshold is partially different from or completely different from the ith one of the eighth threshold to the fourteenth threshold.

[00232] Optionally, the first communication apparatus may determine values of the eighth threshold to the fourteenth threshold in a plurality of manners. For example, one or more of the eighth threshold to the fourteenth threshold may be preconfigured. For another example, one or more of the eighth threshold to the fourteenth threshold may be configured by another device (for example, the second communication apparatus, a network device, or a server) for the first communication apparatus.

[00233] When the first condition is met, the first communication apparatus may determine that current running of the second AI model may be abnormal. Therefore, the first communication

apparatus may send the second information, so that the second communication apparatus can obtain the second information obtained by performing noise perturbation on the first parameter, and update the second AI model based on the second information. In this manner, reduction or even disappear of plasticity in the second AI model during long-term continual learning can be alleviated or avoided, thereby improving adaptability of continual learning and enhancing robustness of continual learning.

[00234] Based on the solution shown in FIG. 5, in S502, the first communication apparatus may send, based on the first information, the second information obtained by performing noise perturbation on the first parameter, so that a receiver of the second information (for example, the second communication apparatus) can obtain the second information obtained by performing noise perturbation on the first parameter, and update the second AI model based on the second information. In comparison with a manner of directly sending the first parameter without performing noise perturbation, in this manner, reduction or even disappear of plasticity in the second AI model during long-term continual learning can be alleviated or avoided, thereby improving adaptability of continual learning and enhancing robustness of continual learning.

[00235] It should be noted that the method shown in FIG. 5 and the method shown in FIG. 3 may be combined with each other. For example, after S301 shown in FIG. 3, S502 shown in FIG. 5 may be further performed. An execution sequence of S302 and S502 is not limited. To be specific, S302 may be performed before S502, or S502 may be performed before S302.

[00236] In a possible implementation of the method shown in FIG. 5, before S502, the method further includes: The first communication apparatus receives indication information indicating a second parameter, where the second parameter is used to process the first parameter to obtain the second information. Specifically, the first communication apparatus may further determine the second parameter based on the received indication information, and process the first parameter based on the second parameter to obtain the second information. In other words, the first communication apparatus may determine the second parameter based on an indication of another communication apparatus (for example, the receiver of the second information), so that the receiver of the second information can obtain the corresponding second information based on the specified second parameter.

[00237] For example, the second communication apparatus may send, to the first communication apparatus based on the first information, the indication information indicating the second parameter. In other words, the second communication apparatus may send, when determining that the first condition is met, the indication information indicating the second parameter, so that the first communication apparatus obtains the second parameter and performs S502.

[00238] Optionally, the indication information indicating the second parameter may include at least one of the second parameter or an index of the second parameter (for an implementation of the index, refer to the manner shown in Table 2 or Table 3).

[00239] It should be understood that the second parameter may be used to perform noise perturbation on the first parameter to obtain the second information. Correspondingly, the second parameter may be understood as a noise perturbation parameter, a noise addition parameter, a perturbation parameter, or the like.

[00240] Optionally, before the first communication apparatus receives the indication information indicating the second parameter, the method further includes: The first communication apparatus sends request information for requesting the second parameter. Therefore, the first communication apparatus may determine the second parameter through an interaction process of the request information and the indication information. For example, the first communication apparatus may send, based on the first information, the request information for requesting the second parameter. In other words, the first communication apparatus may send, when determining that the first condition is met, the request information for requesting the second parameter, to obtain the second parameter and perform S502.

[00241] Optionally, the second parameter is a preconfigured parameter.

[00242] Refer to FIG. 6. An embodiment of this application provides a communication apparatus 600. The communication apparatus 600 can implement functions of the first communication apparatus (or the second communication apparatus) in the foregoing method embodiments, and therefore can achieve beneficial effects of the foregoing method embodiments. In this embodiment of this application, the communication apparatus 600 may be the first communication apparatus (or the second communication apparatus), or may be an integrated circuit, a component, or the like inside the first communication apparatus (or the second communication apparatus), for example, a chip, a baseband chip, a modem chip, a SoC chip including a modem core, a system-in-package (system-in-package, SIP) chip, a communication module, a system on chip, or a processor.

[00243] It should be noted that a transceiver unit 602 may include a sending unit and a receiving unit that are respectively configured to perform sending and receiving.

[00244] In a possible implementation, when the communication apparatus 600 is configured to perform the method performed by the first communication apparatus in FIG. 3 and the related embodiment, the communication apparatus 600 includes a processing unit 601. The processing unit 601 is configured to obtain first information, where the first information indicates performance monitoring information of at least one AI model, and the at least one AI model is associated with a first AI model deployed in the first communication apparatus. The processing unit 601 is further

configured to perform noise perturbation on a model parameter of the first AI model based on the first information, to obtain an updated model parameter of the first AI model.

[00245] In a possible implementation, when the communication apparatus 600 is configured to perform the method performed by the first communication apparatus in FIG. 5 and the related embodiment, the communication apparatus 600 includes the processing unit 601. The processing unit 601 is configured to obtain first information, where the first information indicates performance monitoring information of at least one AI model, and the at least one AI model is associated with a second AI model deployed in the second communication apparatus. The transceiver unit 602 is configured to send second information based on the first information, where the second information is obtained by performing noise perturbation on a first parameter, and the second information is used to update the second AI model.

[00246] In a possible implementation, when the communication apparatus 600 is configured to perform the method performed by the second communication apparatus in FIG. 5 and the related embodiment, the communication apparatus 600 includes the processing unit 601. The transceiver unit 602 is configured to receive second information, where the second information is obtained by performing noise perturbation on a first parameter, and the second information is used to update a second AI model deployed in the second communication apparatus. The processing unit 601 is configured to update the second AI model based on the second information.

[00247] In a possible design, when the communication apparatus 600 is a terminal device or a communication module in a terminal, functions of the processing unit 601 may be implemented by one or more processors. Specifically, the processor may include a modem chip, or a SoC chip including a modem core, or a SIP chip. Functions of the transceiver unit 602 may be implemented by a transceiver circuit.

[00248] In a possible design, when the communication apparatus 600 is a circuit or chip for implementing a communication function in a terminal, for example, a modem chip, a SoC chip including a modem core, or a SIP chip, functions of the processing unit 601 may be implemented by a circuit system that includes one or more processors or processor cores in the chip. Functions of the transceiver unit 602 may be implemented by an interface circuit or a data transceiver circuit in the chip.

[00249] It should be noted that, for content such as an information execution process of the unit of the communication apparatus 600, refer to the descriptions in the foregoing method embodiment in this application. Details are not described herein.

[00250] FIG. 7 is a diagram of another structure of a communication apparatus 700 according to this application. The communication apparatus 700 includes a logic circuit 701 and an input/output interface 702. The communication apparatus 700 may be a chip or an integrated

circuit.

[00251] The transceiver unit 602 shown in FIG. 6 may include a communication interface. The communication interface may include the input/output interface 702 in FIG. 7, and the input/output interface 702 may include an input interface and an output interface. Alternatively, the communication interface may be a transceiver circuit, and the transceiver circuit may include an input interface circuit and an output interface circuit.

[00252] In a possible implementation, when the communication apparatus 700 is configured to perform the method performed by the first communication apparatus in FIG. 3 and the related embodiment, the logic circuit 701 is configured to obtain first information, where the first information indicates performance monitoring information of at least one AI model, and the at least one AI model is associated with a first AI model deployed in the first communication apparatus. The logic circuit 701 is further configured to perform noise perturbation on a model parameter of the first AI model based on the first information, to obtain an updated model parameter of the first AI model.

[00253] In a possible implementation, when the communication apparatus 700 is configured to perform the method performed by the first communication apparatus in FIG. 5 and the related embodiment, the logic circuit 701 is configured to obtain first information, where the first information indicates performance monitoring information of at least one AI model, and the at least one AI model is associated with a second AI model deployed in a second communication apparatus. The input/output interface 702 is configured to send second information based on the first information, where the second information is obtained by performing noise perturbation on a first parameter, and the second information is used to update the second AI model.

[00254] In a possible implementation, when the communication apparatus 700 is configured to perform the method performed by the second communication apparatus in FIG. 5 and the related embodiment, the input/output interface 702 is configured to receive second information, where the second information is obtained by performing noise perturbation on a first parameter, and the second information is used to update a second AI model deployed in the second communication apparatus. The logic circuit 701 is configured to update the second AI model based on the second information.

[00255] The logic circuit 701 and the input/output interface 702 may further perform other steps performed by the first communication apparatus or the second communication apparatus in any embodiment, and achieve corresponding beneficial effects. Details are not described herein again.

[00256] In a possible implementation, the processing unit 601 shown in FIG. 6 may be the logic circuit 701 in FIG. 7.

[00257] Optionally, the logic circuit 701 may be a processing apparatus, and some or all

functions of the processing apparatus may be implemented by software. Some or all functions of the processing apparatus may be implemented by software.

[00258] Optionally, the processing apparatus may include a memory and a processor. The memory is configured to store a computer program, and the processor reads and executes the computer program stored in the memory, to perform corresponding processing and/or steps in any method embodiment.

[00259] Optionally, the processing apparatus may include only the processor. The memory configured to store the computer program is located outside the processing apparatus, and the processor is connected to the memory through a circuit/wire, to read and execute the computer program stored in the memory. The memory and the processor may be integrated together, or may be physically independent of each other.

[00260] Optionally, the processing apparatus may be one or more chips or one or more integrated circuits. For example, the processing apparatus may be one or more field-programmable gate arrays (field-programmable gate arrays, FPGAs), an application-specific integrated circuit (application-specific integrated circuit, ASIC), a system on chip (system on chip, SoC), a central processing unit (central processing unit, CPU), a network processor (network processor, NP), a digital signal processing circuit (digital signal processor, DSP), a microcontroller unit (microcontroller unit, MCU), a programmable logic controller (programmable logic device, PLD) or another integrated chip, or any combination of the foregoing chips or processors.

[00261] FIG. 8 shows a communication apparatus 800 in the foregoing embodiments according to an embodiment of this application. The communication apparatus 800 may be specifically a communication apparatus serving as a terminal device in the foregoing embodiments. An example shown in FIG. 8 is that the terminal device is implemented as the terminal device (or a component in the terminal device).

[00262] In a possible diagram of a logical structure of the communication apparatus 800, the communication apparatus 800 may include but is not limited to at least one processor 801 and a communication port 802.

[00263] The transceiver unit 602 shown in FIG. 6 may include a communication interface. The communication interface may include the communication port 802 in FIG. 8, and the communication port 802 may include an input interface and an output interface. Alternatively, the communication port 802 may be a transceiver circuit, and the transceiver circuit may include an input interface circuit and an output interface circuit.

[00264] Further, optionally, the communication apparatus 800 may further include at least one of a memory 803 and a bus 804. In this embodiment of this application, the at least one processor 801 is configured to control an action of the communication apparatus 800.

[00265] In addition, the processor 801 may be a central processing unit, a general-purpose processor, a digital signal processor, an application-specific integrated circuit, a field programmable gate array or another programmable logic device, a transistor logic device, a hardware component, or any combination thereof. The processor may implement or execute various example logical blocks, modules, and circuits described with reference to content disclosed in this application. Alternatively, the processor may be a combination of processors implementing a computational function, for example, a combination of one or more microprocessors, or a combination of a digital signal processor and a microprocessor. It may be clearly understood by a person skilled in the art that, for the purpose of convenient and brief description, for a detailed operating process of the foregoing system, apparatus, and unit, refer to a corresponding process in the foregoing method embodiment. Details are not described herein again.

[00266] It should be noted that the communication apparatus 800 shown in FIG. 8 may be specifically configured to implement steps implemented by the terminal device in the foregoing method embodiments, and achieve technical effects corresponding to the terminal device. For a specific implementation of the communication apparatus shown in FIG. 8, refer to the descriptions in the foregoing method embodiments. Details are not described herein again.

[00267] FIG. 9 is a diagram of a structure of a communication apparatus 900 in the foregoing embodiments according to an embodiment of this application. The communication apparatus 900 may be specifically a communication apparatus serving as the network device in the foregoing embodiments. An example shown in FIG. 9 is that the network device is implemented as the network device (or a component in the network device). For a structure of the communication apparatus, refer to the structure shown in FIG. 9.

[00268] The communication apparatus 900 includes at least one processor 911 and at least one network interface 914. Further, optionally, the communication apparatus further includes at least one memory 912, at least one transceiver 913, and one or more antennas 915. The processor 911, the memory 912, the transceiver 913, and the network interface 914 are connected, for example, connected through a bus. In this embodiment of this application, the connection may include various interfaces, transmission lines, buses, or the like. This is not limited in embodiments. The antenna 915 is connected to the transceiver 913. The network interface 914 is configured to cause the communication apparatus to communicate with another communication device through a communication link. For example, the network interface 914 may include a network interface between the communication apparatus and a core network device, for example, an S1 interface. The network interface may include a network interface between the communication apparatus and another communication apparatus (for example, another network device or a core network device), for example, an X2 or Xn interface.

[00269] The transceiver unit 602 shown in FIG. 6 may include a communication interface. The communication interface may include the network interface 914 in FIG. 9, and the network interface 914 may include an input interface and an output interface. Alternatively, the network interface 914 may be a transceiver circuit, and the transceiver circuit may include an input interface circuit and an output interface circuit.

[00270] The processor 911 is mainly configured to: process a communication protocol and communication data; control the entire communication apparatus, execute a software program, and process data of the software program. For example, the processor 911 is configured to support the communication apparatus in performing actions described in embodiments. The communication apparatus may include a baseband processor and a central processing unit. The baseband processor is mainly configured to process the communication protocol and the communication data. The central processing unit is mainly configured to: control an entire terminal device, execute the software program, and process the data of the software program. The processor 911 in FIG. 9 may integrate functions of the baseband processor and the central processing unit. A person skilled in the art may understand that the baseband processor and the central processing unit may alternatively be processors independent of each other, and are interconnected through a bus or by using another technology. A person skilled in the art may understand that the terminal device may include a plurality of baseband processors to adapt to different network standards, and the terminal device may include a plurality of central processing units to enhance processing capabilities of the terminal device, and components of the terminal device may be connected through various buses. The baseband processor may also be referred to as a baseband processing circuit or a baseband processing chip. The central processing unit may also be referred to as a central processing circuit or a central processing chip. A function of processing the communication protocol and the communication data may be built in the processor, or may be stored in the memory in a form of software program, and the processor executes the software program to implement a baseband processing function.

[00271] The memory is mainly configured to store the software program and data. The memory 912 may exist independently, and is connected to the processor 911. Optionally, the memory 912 may be integrated with the processor 911, for example, integrated into a chip. The memory 912 can store program code for executing the technical solutions in embodiments of this application, and the processor 911 controls the execution. Various types of executed computer program code may also be considered as drivers of the processor 911.

[00272] FIG. 9 shows only one memory and one processor. In an actual terminal device, there may be a plurality of processors and a plurality of memories. The memory may also be referred to as a storage medium, a storage device, or the like. The memory may be a storage element on a

same chip as the processor, namely, an on-chip storage element, or may be an independent storage element. This is not limited in embodiments of this application.

[00273] The transceiver 913 may be configured to support receiving or sending of a radio frequency signal between the communication apparatus and the terminal, and the transceiver 913 may be connected to the antennas 915. The transceiver 913 includes a transmitter Tx and a receiver Rx. Specifically, the one or more antennas 915 may receive a radio frequency signal. The receiver Rx of the transceiver 913 is configured to: receive the radio frequency signal from the antenna, convert the radio frequency signal into a digital baseband signal or a digital intermediate frequency signal, and provide the digital baseband signal or the digital intermediate frequency signal for the processor 911, so that the processor 911 further processes, for example, demodulates or decodes, the digital baseband signal or the digital intermediate frequency signal. In addition, the transmitter Tx of the transceiver 913 is further configured to: receive a modulated digital baseband signal or a modulated digital intermediate frequency signal from the processor 911, convert the modulated digital baseband signal or the modulated digital intermediate frequency signal into a radio frequency signal, and send the radio frequency signal through the one or more antennas 915. Specifically, the receiver Rx may selectively perform one-level or multi-level down-conversion mixing and analog-to-digital conversion on the radio frequency signal, to obtain the digital baseband signal or the digital intermediate frequency signal. A sequence of the down-conversion mixing and the analog-to-digital conversion may be adjusted. The transmitter Tx may selectively perform one-level or multi-level up-conversion mixing and digital-to-analog conversion on the modulated digital baseband signal or digital intermediate frequency signal, to obtain the radio frequency signal. A sequence of the up-conversion mixing and the digital-to-analog conversion may be adjusted. The digital baseband signal and the digital intermediate frequency signal may be collectively referred to as a digital signal.

[00274] The transceiver 913 may also be referred to as a transceiver unit, a transceiver machine, a transceiver apparatus, or the like. Optionally, a component configured to implement a receiving function in the transceiver unit may be considered as a receiving unit, and a component configured to implement a sending function in the transceiver unit may be considered as a sending unit. In other words, the transceiver unit includes a receiving unit and a sending unit. The receiving unit may also be referred to as a receiver, an input port, a receiver circuit, or the like. The sending unit may be referred to as a transmitter, a transmitter machine, a transmitter circuit, or the like.

[00275] It should be noted that the communication apparatus 900 shown in FIG. 9 may be specifically configured to implement steps implemented by the network device in the foregoing method embodiments, and achieve technical effects corresponding to the network device. For a specific implementation of the communication apparatus 900 shown in FIG. 9, refer to the

descriptions in the foregoing method embodiments. Details are not described herein again.

[00276] FIG. 10 is a diagram of a structure of a communication apparatus in the foregoing embodiments according to an embodiment of this application.

[00277] It may be understood that the communication apparatus 100 includes, for example, modules, units, elements, circuits, or interfaces, which are appropriately configured together to perform the technical solutions provided in this application. The communication apparatus 100 may be the terminal device or the network device described above, or may be a component (for example, a chip) in these devices, to implement the methods described in the following method embodiments. The communication apparatus 100 includes one or more processors 101. The processor 101 may be a general-purpose processor, a dedicated processor, or the like. For example, the processor 101 may be a baseband processor or a central processing unit. The baseband processor may be configured to process a communication protocol and communication data. The central processing unit may be configured to: control the communication apparatus (for example, a RAN node, a terminal, or a chip), execute a software program, and process data of the software program.

[00278] Optionally, in a design, the processor 101 may include a program 103 (which may also be referred to as code or instructions sometimes). The program 103 may be run on the processor 101, to cause the communication apparatus 100 to perform the methods described in the following embodiments. In another possible design, the communication apparatus 100 includes a circuit (not shown in FIG. 10).

[00279] Optionally, the communication apparatus 100 may include one or more memories 102, storing a program 104 (which may also be referred to as code or instructions sometimes). The program 104 may be run on the processor 101, to cause the communication apparatus 100 to perform the methods described in the foregoing method embodiments.

[00280] Optionally, the processor 101 and/or the memory 102 may include AI modules 107 and 108, and the AI modules are configured to implement an AI-related function. The AI module may be implemented by software, hardware, or a combination of software and hardware. For example, the AI module may include a radio intelligence control (radio intelligence control, RIC) module. For example, the AI module may be a near-real-time RIC or a non-real-time RIC.

[00281] Optionally, the processor 101 and/or the memory 102 may further store data. The processor and the memory may be separately disposed, or may be integrated together.

[00282] Optionally, the communication apparatus 100 may further include a transceiver 105 and/or an antenna 106. The processor 101 may also be referred to as a processing unit sometimes, and control the communication apparatus (for example, the RAN node or the terminal). The transceiver 105 may also be referred to as a transceiver unit, a transceiver machine, a transceiver

circuit, or the like sometimes, and is configured to implement a transceiver function of the communication apparatus through the antenna 106.

[00283] The processing unit 601 shown in FIG. 6 may be the processor 101. The transceiver unit 602 shown in FIG. 6 may be a communication interface. The communication interface may be the transceiver 105 in FIG. 10. The transceiver 105 may include an input interface and an output interface. Alternatively, the transceiver 105 may be a transceiver circuit, and the transceiver circuit may include an input interface circuit and an output interface circuit.

[00284] An embodiment of this application further provides a computer-readable storage medium. The storage medium is configured to store one or more computer-executable instructions. When the computer-executable instructions are executed by a processor, the processor performs the method in the possible implementations of the first communication apparatus or the second communication apparatus in the foregoing embodiments.

[00285] An embodiment of this application further provides a computer program product (or referred to as a computer program). When the computer program product is executed by a processor, the processor performs the method in the possible implementations of the first communication apparatus or the second communication apparatus.

[00286] An embodiment of this application further provides a system on chip. The system on chip includes at least one processor, configured to support a communication apparatus in implementing functions in the possible implementations of the communication apparatus. Optionally, the system on chip further includes an interface circuit, and the interface circuit provides program instructions and/or data for the at least one processor. In a possible design, the system on chip may further include a memory, and the memory is configured to store program instructions and data that are necessary for the communication apparatus. The system on chip may include a chip, or may include a chip and another discrete component. The communication apparatus may be specifically the first communication apparatus or the second communication apparatus in the foregoing method embodiments.

[00287] An embodiment of this application further provides a communication system. An architecture of the network system includes the first communication apparatus and/or the second communication apparatus in any one of the foregoing embodiments.

[00288] 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 apparatus embodiments described above are merely examples. For example, the unit division is merely logical function division and may be other division during 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. Whether a function is performed by hardware or software depends on particular applications and design constraints 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.

[00289] The units described as separate parts may or may not be physically separate, and parts displayed as units may or may not be physical units, may be located in one position, or may be distributed on a plurality of network units. Some or all of the units may be selected based on actual requirements to achieve the objectives of the solutions of embodiments.

[00290] In addition, functional units in embodiments of this application may be integrated into one processing unit, each of the units may exist alone physically, or two or more units may be integrated into one unit. The integrated unit may be implemented in a form of hardware, or may be implemented in a form of software functional unit. When the integrated unit is implemented in the form of software functional unit and sold or used as an independent product, the integrated unit may be stored in a computer-readable storage medium. Based on such understanding, a part of the technical solutions of this application that essentially makes contributions, or all or a part of the technical solutions, may be embodied in the form of software product. The computer software product is stored in a storage medium and includes several instructions for instructing a computer device (which may be a personal computer, a server, a network device, or the like) to perform all or some of the steps of the methods described in embodiments of this application. The foregoing storage medium includes any medium that can store program code, such as a USB flash drive, a removable hard disk, a read-only memory (read-only memory, ROM), a random access memory (random access memory, RAM), a magnetic disk, or an optical disc.

CLAIMS

1. A communication method, wherein the method is applied to a first communication apparatus, and the method comprises:

obtaining first information, wherein the first information indicates performance monitoring information of at least one artificial intelligence AI model, and the at least one AI model is associated with a first AI model deployed in the first communication apparatus; and

performing noise perturbation on a model parameter of the first AI model based on the first information, to obtain an updated model parameter of the first AI model.

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

sending second information based on the first information, wherein the second information is obtained by performing noise perturbation on a first parameter, the second information is used to update a second AI model, and the second AI model is deployed in a second communication apparatus.

3. The method according to claim 2, wherein the second information is obtained by processing a quantization result of the first parameter, or the second information is obtained by processing a transmission parameter of the first parameter.

4. The method according to claim 2 or 3, wherein the first parameter comprises one or more of the following:

an updated model parameter of the second AI model, gradient information used to update the second AI model, intermediate feature information used to update the second AI model, intermediate gradient information used to update the second AI model, an inference result of the first AI model, distillation loss information corresponding to the second AI model, or task loss information of the second AI model.

5. The method according to any one of claims 2 to 4, wherein sending the second information based on the first information comprises:

sending the second information based on the first information when a first condition is met, wherein

the first condition comprises at least one of the following:

training accuracy of the at least one AI model is lower than a first threshold;

test accuracy of the at least one AI model is lower than a second threshold;

system performance of an AI model system in which the at least one AI model is located is lower than a third threshold;

system performance of a communication system in which a communication apparatus deployed with the at least one AI model is located is lower than a fourth threshold;

a neural network input-output distribution shift of the at least one AI model is greater than a fifth threshold;

a neural network weight distribution of the at least one AI model is greater than a sixth threshold; or

a neural network weight variation of the at least one AI model is lower than a seventh threshold.

6. The method according to any one of claims 2 to 5, further comprising:

receiving indication information indicating the second parameter, wherein the second parameter is used to process the first parameter to obtain the second information.

7. The method according to claim 6, further comprising:

sending request information for requesting the second parameter.

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

receiving indication information indicating a third parameter, wherein the third parameter is used to process the model parameter of the first AI model to obtain the updated model parameter of the first AI model.

9. The method according to claim 8, further comprising:

sending request information for requesting the third parameter.

10. The method according to any one of claims 1 to 9, wherein performing noise perturbation on the model parameter of the first AI model based on the first information, to obtain the updated model parameter of the first AI model comprises:

when a second condition is met, performing noise perturbation on the model parameter of the first AI model based on the first information, to obtain the updated model parameter of the first AI model, wherein

the second condition comprises at least one of the following:

the training accuracy of the at least one AI model is lower than an eighth threshold;

the test accuracy of the at least one AI model is lower than a ninth threshold;

the system performance of the AI model system in which the at least one AI model is located is lower than a tenth threshold;

the system performance of the communication system in which the communication apparatus deployed with the at least one AI model is located is lower than an eleventh threshold;

the neural network input-output distribution shift of the at least one AI model is greater than a twelfth threshold;

the neural network weight distribution of the at least one AI model is greater than a thirteenth threshold; or

the neural network weight variation of the at least one AI model is lower than a fourteenth

threshold.

11. The method according to any one of claims 1 to 10, wherein obtaining the first information comprises:

receiving the first information; or

obtaining a measurement parameter of the first AI model and/or a measurement parameter of the second AI model, and determining the first information based on the measurement parameter of the first AI model.

12. The method according to any one of claims 1 to 11, wherein the at least one AI model comprises the first AI model and/or the second AI model, the first AI model is associated with the second AI model, and the second AI model is deployed in the second communication apparatus.

13. A communication method, applied to a first communication apparatus, wherein the method comprises:

obtaining first information, wherein the first information indicates performance monitoring information of at least one artificial intelligence AI model, and the at least one AI model is associated with a second AI model deployed in a second communication apparatus; and

sending second information based on the first information, wherein the second information is obtained by performing noise perturbation on a first parameter, and the second information is used to update the second AI model.

14. The method according to claim 13, wherein the second information is obtained by processing a quantization result of the first parameter, or the second information is obtained by processing a transmission parameter of the first parameter.

15. The method according to claim 13 or 14, wherein the first parameter comprises one or more of the following:

an updated model parameter of the second AI model, gradient information used to update the second AI model, intermediate feature information used to update the second AI model, intermediate gradient information used to update the second AI model, an inference result of the first AI model, distillation loss information corresponding to the second AI model, or task loss information of the second AI model.

16. The method according to any one of claims 13 to 15, wherein sending the second information based on the first information comprises: sending the second information based on the first information when a first condition is met, wherein

the first condition comprises at least one of the following:

training accuracy of the at least one AI model is lower than a first threshold;

test accuracy of the at least one AI model is lower than a second threshold;

system performance of an AI model system in which the at least one AI model is located is

lower than a third threshold;

system performance of a communication system in which a communication apparatus deployed with the at least one AI model is located is lower than a fourth threshold;

a neural network input-output distribution shift of the at least one AI model is greater than a fifth threshold;

a neural network weight distribution of the at least one AI model is greater than a sixth threshold; or

a neural network weight variation of the at least one AI model is lower than a seventh threshold.

17. The method according to any one of claims 13 to 16, further comprising:

receiving indication information indicating the second parameter, wherein the second parameter is used to process the first parameter to obtain the second information.

18. The method according to claim 17, further comprising:

sending request information for requesting the second parameter.

19. The method according to any one of claims 13 to 18, wherein obtaining the first information comprises:

receiving the first information; or

obtaining a measurement parameter of the first AI model and/or a measurement parameter of the second AI model, and determining the first information based on the measurement parameter of the first AI model.

20. The method according to any one of claims 13 to 19, wherein the at least one AI model comprises the first AI model and/or the second AI model, the first AI model is associated with the second AI model, and the second AI model is deployed in the second communication apparatus.

21. A communication method, applied to a second communication apparatus, wherein the method comprises:

receiving second information, wherein the second information is obtained by performing noise perturbation on a first parameter, and the second information is used to update a second artificial intelligence AI model deployed in the second communication apparatus; and

updating the second AI model based on the second information.

22. The method according to claim 21, wherein the second information is obtained by processing a quantization result of the first parameter, or the second information is obtained by processing a transmission parameter of the first parameter.

23. The method according to claim 21 or 22, wherein the first parameter comprises one or more of the following:

an updated model parameter of the second AI model, gradient information used to update the

second AI model, intermediate feature information used to update the second AI model, intermediate gradient information used to update the second AI model, an inference result of the first AI model, distillation loss information corresponding to the second AI model, or task loss information of the second AI model.

24. A first communication apparatus, comprising a processing unit, wherein

the processing unit is configured to: obtain first information, wherein the first information indicates performance monitoring information of at least one artificial intelligence AI model, and the at least one AI model is associated with a first AI model deployed in the first communication apparatus; and

perform noise perturbation on a model parameter of the first AI model based on the first information, to obtain an updated model parameter of the first AI model.

25. The apparatus according to claim 24, wherein the apparatus further comprises a transceiver unit; and

the transceiver unit is configured to send second information based on the first information, wherein the second information is obtained by performing noise perturbation on a first parameter, the second information is used to update a second AI model, and the second AI model is deployed in a second communication apparatus.

26. The apparatus according to claim 25, wherein the second information is obtained by processing a quantization result of the first parameter, or the second information is obtained by processing a transmission parameter of the first parameter.

27. The apparatus according to claim 25 or 26, wherein the first parameter comprises one or more of the following:

an updated model parameter of the second AI model, gradient information used to update the second AI model, intermediate feature information used to update the second AI model, intermediate gradient information used to update the second AI model, an inference result of the first AI model, distillation loss information corresponding to the second AI model, or task loss information of the second AI model.

28. The apparatus according to any one of claims 25 to 27, wherein

the transceiver unit is further configured to:

send the second information based on the first information when a first condition is met, wherein

the first condition comprises at least one of the following:

training accuracy of the at least one AI model is lower than a first threshold;

test accuracy of the at least one AI model is lower than a second threshold;

system performance of an AI model system in which the at least one AI model is located is

lower than a third threshold;

system performance of a communication system in which a communication apparatus deployed with the at least one AI model is located is lower than a fourth threshold;

a neural network input-output distribution shift of the at least one AI model is greater than a fifth threshold;

a neural network weight distribution of the at least one AI model is greater than a sixth threshold; or

a neural network weight variation of the at least one AI model is lower than a seventh threshold.

29. The apparatus according to any one of claims 25 to 28, wherein

the transceiver unit is further configured to receive indication information indicating the second parameter, wherein the second parameter is used to process the first parameter to obtain the second information.

30. The apparatus according to claim 29, wherein

the transceiver unit is further configured to send request information for requesting the second parameter.

31. The apparatus according to any one of claims 24 to 30, wherein

the transceiver unit is further configured to receive indication information indicating a third parameter, wherein the third parameter is used to process the model parameter of the first AI model to obtain the updated model parameter of the first AI model.

32. The apparatus according to claim 31, wherein

the transceiver unit is further configured to send request information for requesting the third parameter.

33. The apparatus according to any one of claims 24 to 32, wherein

the processing unit is further configured to:

when a second condition is met, perform noise perturbation on the model parameter of the first AI model based on the first information, to obtain the updated model parameter of the first AI model, wherein

the second condition comprises at least one of the following:

the training accuracy of the at least one AI model is lower than an eighth threshold;

the test accuracy of the at least one AI model is lower than a ninth threshold;

the system performance of the AI model system in which the at least one AI model is located is lower than a tenth threshold;

the system performance of the communication system in which the communication apparatus deployed with the at least one AI model is located is lower than an eleventh threshold;

the neural network input-output distribution shift of the at least one AI model is greater than a twelfth threshold;

the neural network weight distribution of the at least one AI model is greater than a thirteenth threshold; or

the neural network weight variation of the at least one AI model is lower than a fourteenth threshold.

34. The apparatus according to any one of claims 24 to 33, wherein

the processing unit is further configured to: receive the first information; or

obtain a measurement parameter of the first AI model and/or a measurement parameter of the second AI model, and determine the first information based on the measurement parameter of the first AI model.

35. The apparatus according to any one of claims 24 to 34, wherein the at least one AI model comprises the first AI model and/or the second AI model, the first AI model is associated with the second AI model, and the second AI model is deployed in the second communication apparatus.

36. A first communication apparatus, comprising a processing unit, wherein

the processing unit is configured to: obtain first information, wherein the first information indicates performance monitoring information of at least one artificial intelligence AI model, and the at least one AI model is associated with a second AI model deployed in a second communication apparatus; and

send second information based on the first information, wherein the second information is obtained by performing noise perturbation on a first parameter, and the second information is used to update the second AI model.

37. The apparatus according to claim 36, wherein the second information is obtained by processing a quantization result of the first parameter, or the second information is obtained by processing a transmission parameter of the first parameter.

38. The apparatus according to claim 36 or 37, wherein the first parameter comprises one or more of the following:

an updated model parameter of the second AI model, gradient information used to update the second AI model, intermediate feature information used to update the second AI model, intermediate gradient information used to update the second AI model, an inference result of the first AI model, distillation loss information corresponding to the second AI model, or task loss information of the second AI model.

39. The apparatus according to any one of claims 36 to 38, wherein the transceiver unit is further configured to:

send the second information based on the first information when a first condition is met,

wherein

the first condition comprises at least one of the following:

training accuracy of the at least one AI model is lower than a first threshold;

test accuracy of the at least one AI model is lower than a second threshold;

system performance of an AI model system in which the at least one AI model is located is lower than a third threshold;

system performance of a communication system in which a communication apparatus deployed with the at least one AI model is located is lower than a fourth threshold;

a neural network input-output distribution shift of the at least one AI model is greater than a fifth threshold;

a neural network weight distribution of the at least one AI model is greater than a sixth threshold; or

a neural network weight variation of the at least one AI model is lower than a seventh threshold.

40. The apparatus according to any one of claims 36 to 39, wherein

the transceiver unit is further configured to receive indication information indicating the second parameter, wherein the second parameter is used to process the first parameter to obtain the second information.

41. The apparatus according to claim 40, wherein

the transceiver unit is further configured to send request information for requesting the second parameter.

42. The apparatus according to any one of claims 36 to 41, wherein

the processing unit is further configured to: receive the first information; or

obtain a measurement parameter of the first AI model and/or a measurement parameter of the second AI model, and determine the first information based on the measurement parameter of the first AI model.

43. The apparatus according to any one of claims 36 to 42, wherein the at least one AI model comprises the first AI model and/or the second AI model, the first AI model is associated with the second AI model, and the second AI model is deployed in the second communication apparatus.

44. A second communication apparatus, comprising a processing unit and a transceiver unit, wherein

the transceiver unit is configured to receive second information, wherein the second information is obtained by performing noise perturbation on a first parameter, and the second information is used to update a second artificial intelligence AI model deployed in the second communication apparatus; and

the processing unit is configured to update the second AI model based on the second information.

45. The apparatus according to claim 44, wherein the second information is obtained by processing a quantization result of the first parameter, or the second information is obtained by processing a transmission parameter of the first parameter.

46. The apparatus according to claim 44 or 45, wherein the first parameter comprises one or more of the following:

an updated model parameter of the second AI model, gradient information used to update the second AI model, intermediate feature information used to update the second AI model, intermediate gradient information used to update the second AI model, an inference result of the first AI model, distillation loss information corresponding to the second AI model, or task loss information of the second AI model.

47. A first communication apparatus, comprising a processor, wherein

the processor is configured to: obtain first information, wherein the first information indicates performance monitoring information of at least one artificial intelligence AI model, and the at least one AI model is associated with a first AI model deployed in the first communication apparatus; and

perform noise perturbation on a model parameter of the first AI model based on the first information, to obtain an updated model parameter of the first AI model.

48. The apparatus according to claim 47, wherein the apparatus further comprises a communication port; and

the communication port is further configured to send second information based on the first information, wherein the second information is obtained by performing noise perturbation on a first parameter, the second information is used to update a second AI model, and the second AI model is deployed in a second communication apparatus.

49. The apparatus according to claim 48, wherein the second information is obtained by processing a quantization result of the first parameter, or the second information is obtained by processing a transmission parameter of the first parameter.

50. The apparatus according to claim 48 or 49, wherein the first parameter comprises one or more of the following:

an updated model parameter of the second AI model, gradient information used to update the second AI model, intermediate feature information used to update the second AI model, intermediate gradient information used to update the second AI model, an inference result of the first AI model, distillation loss information corresponding to the second AI model, or task loss information of the second AI model.

51. The apparatus according to any one of claims 48 to 50, wherein the communication port is further configured to:

send the second information based on the first information when a first condition is met, wherein

the first condition comprises at least one of the following:

training accuracy of the at least one AI model is lower than a first threshold;

test accuracy of the at least one AI model is lower than a second threshold;

system performance of an AI model system in which the at least one AI model is located is lower than a third threshold;

system performance of a communication system in which a communication apparatus deployed with the at least one AI model is located is lower than a fourth threshold;

a neural network input-output distribution shift of the at least one AI model is greater than a fifth threshold;

a neural network weight distribution of the at least one AI model is greater than a sixth threshold; or

a neural network weight variation of the at least one AI model is lower than a seventh threshold.

52. The apparatus according to any one of claims 48 to 51, wherein

the communication port is further configured to receive indication information indicating the second parameter, wherein the second parameter is used to process the first parameter to obtain the second information.

53. The apparatus according to claim 52, wherein

the communication port is further configured to send request information for requesting the second parameter.

54. The apparatus according to any one of claims 47 to 53, wherein

the communication port is further configured to receive indication information indicating a third parameter, wherein the third parameter is used to process the model parameter of the first AI model to obtain the updated model parameter of the first AI model.

55. The apparatus according to claim 54, wherein

the communication port is further configured to send request information for requesting the third parameter.

56. The apparatus according to any one of claims 47 to 55, wherein

the processor is further configured to:

when a second condition is met, perform noise perturbation on the model parameter of the first AI model based on the first information, to obtain the updated model parameter of the first AI

model, wherein

the second condition comprises at least one of the following:

the training accuracy of the at least one AI model is lower than an eighth threshold;

the test accuracy of the at least one AI model is lower than a ninth threshold;

the system performance of the AI model system in which the at least one AI model is located is lower than a tenth threshold;

the system performance of the communication system in which the communication apparatus deployed with the at least one AI model is located is lower than an eleventh threshold;

the neural network input-output distribution shift of the at least one AI model is greater than a twelfth threshold;

the neural network weight distribution of the at least one AI model is greater than a thirteenth threshold; or

the neural network weight variation of the at least one AI model is lower than a fourteenth threshold.

57. The apparatus according to any one of claims 47 to 56, wherein

the processor is further configured to: receive the first information; or

obtain a measurement parameter of the first AI model and/or a measurement parameter of the second AI model, and determine the first information based on the measurement parameter of the first AI model.

58. The apparatus according to any one of claims 47 to 57, wherein the at least one AI model comprises the first AI model and/or the second AI model, the first AI model is associated with the second AI model, and the second AI model is deployed in the second communication apparatus.

59. A first communication apparatus, comprising a processor and a communication port, wherein

the processor is configured to: obtain first information, wherein the first information indicates performance monitoring information of at least one artificial intelligence AI model, and the at least one AI model is associated with a second AI model deployed in a second communication apparatus; and

the communication port is configured to send second information based on the first information, wherein the second information is obtained by performing noise perturbation on a first parameter, and the second information is used to update the second AI model.

60. The apparatus according to claim 59, wherein the second information is obtained by processing a quantization result of the first parameter, or the second information is obtained by processing a transmission parameter of the first parameter.

61. The apparatus according to claim 59 or 60, wherein the first parameter comprises one or

more of the following:

an updated model parameter of the second AI model, gradient information used to update the second AI model, intermediate feature information used to update the second AI model, intermediate gradient information used to update the second AI model, an inference result of the first AI model, distillation loss information corresponding to the second AI model, or task loss information of the second AI model.

62. The apparatus according to any one of claims 59 to 61, wherein

the communication port is further configured to:

send the second information based on the first information when a first condition is met, wherein

the first condition comprises at least one of the following:

training accuracy of the at least one AI model is lower than a first threshold;

test accuracy of the at least one AI model is lower than a second threshold;

system performance of an AI model system in which the at least one AI model is located is lower than a third threshold;

system performance of a communication system in which a communication apparatus deployed with the at least one AI model is located is lower than a fourth threshold;

a neural network input-output distribution shift of the at least one AI model is greater than a fifth threshold;

a neural network weight distribution of the at least one AI model is greater than a sixth threshold; or

a neural network weight variation of the at least one AI model is lower than a seventh threshold.

63. The apparatus according to any one of claims 59 to 62, wherein

the communication port is further configured to receive indication information indicating the second parameter, wherein the second parameter is used to process the first parameter to obtain the second information.

64. The apparatus according to claim 63, wherein

the communication port is further configured to send request information for requesting the second parameter.

65. The apparatus according to any one of claims 59 to 64, wherein

the processor is further configured to: receive the first information; or

obtain a measurement parameter of the first AI model and/or a measurement parameter of the second AI model, and determine the first information based on the measurement parameter of the first AI model.

66. The apparatus according to any one of claims 59 to 65, wherein the at least one AI model comprises the first AI model and/or the second AI model, the first AI model is associated with the second AI model, and the second AI model is deployed in the second communication apparatus.

67. A second communication apparatus, comprising a processor and a communication port, wherein

the communication port is configured to receive second information, wherein the second information is obtained by performing noise perturbation on a first parameter, and the second information is used to update a second artificial intelligence AI model deployed in the second communication apparatus; and

the processor is configured to update the second AI model based on the second information.

68. The apparatus according to claim 67, wherein the second information is obtained by processing a quantization result of the first parameter, or the second information is obtained by processing a transmission parameter of the first parameter.

69. The apparatus according to claim 67 or 68, wherein the first parameter comprises one or more of the following:

an updated model parameter of the second AI model, gradient information used to update the second AI model, intermediate feature information used to update the second AI model, intermediate gradient information used to update the second AI model, an inference result of the first AI model, distillation loss information corresponding to the second AI model, or task loss information of the second AI model.

70. The apparatus according to any one of claims 67 to 69, wherein the apparatus is a chip or a system on chip.

71. A readable storage medium, wherein the storage medium stores a computer program or instructions, and when the computer program or the instructions are executed by a computer, the method according to any one of claims 1 to 23 is implemented.

72. A computer program product, comprising instructions, wherein when the instructions run on a computer, the computer is caused to perform the method according to any one of claims 1 to 23.