(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
19 December 2024 (19.12.2024)

WIPO | PCT

(10) International Publication Number
WO 2024/255037 A1

(51) International Patent Classification:
H04W 24/02 (2009.01)

(21) International Application Number:
PCT/CN2023/124990

(22) International Filing Date:
17 October 2023 (17.10.2023)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
63/507,786 13 June 2023 (13.06.2023) US

(71) Applicant: **HUAWEI TECHNOLOGIES CO., LTD.** [CN/CN]; Huawei Administration Building, Bantian, Longgang, Shenzhen, Guangdong 518129 (CN).

(72) Inventors: **TANG, Hao**; Huawei Administration Building, Bantian, Longgang, Shenzhen, Guangdong 518129 (CN). **GE, Yiqun**; Huawei Administration Building, Bantian, Longgang, Shenzhen, Guangdong 518129 (CN). **MA, Jianglei**; Huawei Administration Building, Bantian, Longgang, Shenzhen, Guangdong 518129 (CN).

(74) Agent: **LONGSUN LEAD IP LTD.**; Room 801-1, Floor 8, Building 3, Block 2, No. 81 Beiqing Road, Haidian District, Beijing 100094 (CN).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report (Art. 21(3))*

(54) Title: COMMUNICATION METHOD AND COMMUNICATION APPARATUS

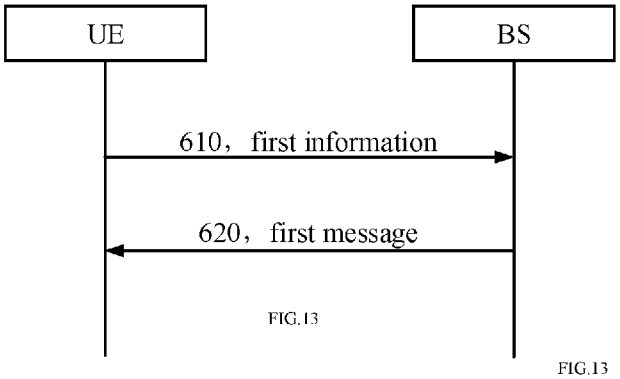


FIG.13

(57) Abstract: Embodiments of the present application provide a communication method and a communication apparatus. The method includes: sending first information related to mutual information of a first artificial intelligence (AI) model and a second AI model, the first AI model and the second AI model constituting a two-sided model; and receiving a first message indicating an AI model related to the first AI model. The AI models at the UE and BS sides constitute a two-sided model, and the UE or the BS can send information related to the mutual information of the AI models to realize interoperability.

WO 2024/255037 A1

COMMUNICATION METHOD AND COMMUNICATION APPARATUS

[0001] The present application is related to, and claims priority to, United States provisional patent application Serial No. 63/507,786, entitled "AI MODEL SELF GENERALIZATION", filed on June 13, 2023.

[0002] The disclosures of the aforementioned applications are hereby incorporated by reference in their entirety.

TECHNICAL FIELD

[0003] Embodiments of the present application relate to the field of communication, and more specifically, to a communication method and a communication apparatus.

BACKGROUND

[0004] AI-based algorithms have been introduced into modern wireless communications to solve some wireless problems such as channel estimation, scheduling, channel state information (CSI) compression (from a user equipment to a base-station), multiple-in multiple-out (MIMO)'s beamforming, positioning, and so on. As data-driven methods, AI-based algorithms inevitably suffer from low generalization. Performance of artificial intelligence (AI) models is only as good as the data they are trained on. Even if the AI model is trained on a large number of data sets, it may also not possess the necessary knowledge to perform effectively in other environments, especially in wireless communication where the channel information is changed rapidly.

[0005] For example, in an auto-encoder model, an encoder is deployed on a user equipment (UE) side and a decoder is deployed on a base station (BS) side. The BS and UE train their models independently and need to align the encoder and decoder. In addition, during inference, the generalization problem at UE or BS needs to be considered. For example, the generalization performance of the UE encoder model is worse than that of the decoder model of the BS, but it is hard to know whether the current encoder model is outdated or not during the inference process.

[0006] Therefore, how to realize the interoperability between the model of BS and the model of UE is an urgent technical problem to be solved.

SUMMARY

[0007] Embodiments of the present application provide a communication method and a communication apparatus. In the technical solutions of the present application, AI models at the UE and BS sides constitute a two-sided model, and the UE or the BS can send information related to the mutual information of the AI models to realize interoperability.

[0008] According to a first aspect, an embodiment of the present application provides a communication method including: sending first information related to mutual information of a first AI model and a second AI model, the first AI model and the second AI model constituting a two-sided model; and receiving a first message indicating an AI model related to the first AI model.

[0009] In the communication method provided by the present application, the AI models at the UE and BS sides constitute a two-sided model, and the UE or the BS can send information related to the mutual information of the AI models to realize interoperability.

[0010] A first AI model is an encoder and a second AI model is a decoder. Alternatively, a first AI model is a decoder and a second AI model is an encoder. The first AI model and the second AI model constitute a two-sided model.

[0011] In one possible scenario, the first AI model is at the UE side and the second AI model is at the BS side.

[0012] In another possible implementation scenario, the first AI model 1 and the second AI model 1 are at the UE side and the first AI model 2 and the second AI model 2 are at the BS side. For example, UE trains its own encoder1 and decoder1, BS trains its own encoder2 and decoder2, and finally aligns the UE's encoder 1 with the BS's decoder 2. In this case, the first information sent by the UE is information related to the mutual information of encoder1 and decoder1.

[0013] In a possible implementation, the first information includes at least one of first mutual information, second mutual information, a first ratio, a first neuron node size, and a first ratio range.

[0014] The first mutual information is an amount of information about an input included in an output of the first AI model. The second mutual information is an amount of information about an output included in an input of the second AI model. The first ratio is a ratio of the second mutual information to the first mutual information. The first neuron node size is configured to indicate an output format of the first AI model or an input format of the second AI model. The first ratio range is a range of multiple first ratios. The ratio range can also be some discrete value, such as a collection of values.

[0015] In a two-sided model, if the encoder is deployed at a UE side and the decoder is deployed at a BS side, X is the input of the encoder, Y is the output of the decoder, and T is the output of the encoder as well as the input of the decoder. T is exchanged through the air interface between the BS and the UE.

[0016] In a two-sided model, if encoder 1 and decoder 1 are deployed at the UE side, X is the input of the encoder 1, Y

is the output of the decoder 1, and T is the output of the encoder 1 and also the input of the decoder 1. T can also be regarded as a latent layer or split point of the joint model of encoder 1 and decoder 1.

[0017] In a two-sided model, if encoder 2 and decoder 2 are deployed at the BS side, X is the input of the encoder 2, Y is the output of the decoder 2, and T is the output of the encoder 2 and also the input of the decoder 2. T can also be regarded as a latent layer or split point of the joint model of encoder 2 and decoder 2.

[0018] For example, let I(X,T) denote the first mutual information and let I(T,Y) denote the second mutual information. The first ratio can be function-1(I(T, Y)) / function-2(I(X, T)), and the function-1 and function-2 can be one of max(), min(), average(), and so on, where max() represents the maximum value, min() represents the minimum value, and average() represents the average value.

[0019] In the communication method provided by the present application, the UE can send at least one of first mutual information, second mutual information, a first ratio, a first neuron node size, and a first ratio range to the BS, and the BS indicates the AI model at the UE side based on the first information, realizing two-sided model interoperability.

[0020] In a possible implementation, the first information includes an index corresponding to first mutual information, an index corresponding to second mutual information, an index corresponding to a first ratio, an index corresponding to a first neuron node size, or an index corresponding to a first ratio range.

[0021] At least one piece of third mutual information, at least one piece of fourth mutual information, at least one second ratio, at least one second neuron node size or at least one second ratio range is predetermined or configured by a network device, where each of third mutual information, fourth mutual information, second ratios, second neuron node sizes or second ratio ranges corresponds to an index. The first mutual information is one of the at least one piece of third mutual information, the second mutual information is one of the at least one piece of fourth mutual information, the first ratio is one of the at least one second ratio, the first neuron node size is one of the at least one second neuron node size, and the first ratio range is one of the at least one second ratio range.

[0022] In the communication method provided by the present application, the UE can send an index corresponding to the first information to the BS, which can reduce the air interface overhead while realizing interoperability.

[0023] In a possible implementation, the sending first information includes: sending the first information when the first information has changed during a time period.

[0024] The reporting of the UE can be event-triggered reporting, e.g., the UE reports when its ratio or ratio range or mutual information changes.

[0025] In the communication method provided by the present application, the UE can send the first information when the first information has changed during a time period, which can reduce the air interface overhead while realizing

interoperability.

[0026] In a possible implementation, the sending first information includes: sending a model switch request.

[0027] In a possible implementation, if the ratio or ratio range or mutual information of the UE is out of range, the UE can send a model switch request to the BS. The BS receives the model switch request and sends the first message indicating the UE to perform the switching of the model.

[0028] In the communication method provided by the present application, the AI models at the UE and BS sides constitute a two-sided model, and the UE or the BS can send information related to the mutual information of the AI models to realize interoperability.

[0029] In a possible implementation, the method further includes: receiving a second message indicating a method for calculating the first information.

[0030] In a possible implementation, the method for calculating the first information is Hilbert-Schmidt independence criterion (HSIC), or a predefined mutual information approximation method.

[0031] The BS can send a mutual information approximation method to the UE, such as Hilbert-Schmidt Independence Criterion (HSIC), or a predefined mutual information approximation method.

[0032] In a possible implementation, the first AI model is a decoder and the second AI model is an encoder; or, the first AI model is an encoder and the second AI model is a decoder.

[0033] In a possible implementation, the first AI model is at a user equipment side and the second AI model is at a network device side; or the first AI model and the second AI model are at a user equipment side.

[0034] In a possible implementation, the method is executed by a user equipment.

[0035] According to a second aspect, an embodiment of the present application provides a communication method including: receiving first information related to mutual information of a first artificial intelligence (AI) model and a second AI model, the first AI model and the second AI model constituting a two-sided model; and sending a first message indicating an AI model related to the first AI model.

[0036] In a possible implementation, the first information includes at least one of first mutual information, second mutual information, a first ratio, a first neuron node size, and a first ratio range.

[0037] In a possible implementation, the first information includes an index corresponding to first mutual information, an index corresponding to second mutual information, an index corresponding to a first ratio, an index corresponding to a first neuron node size, or an index corresponding to a first ratio range.

[0038] In a possible implementation, at least one piece of third mutual information, at least one piece of fourth mutual information, at least one second ratio, at least one second neuron node size or at least one second ratio range is predetermined

or configured by a network device, where each of third mutual information, fourth mutual information, second ratios, second neuron node sizes or second ratio ranges corresponds to an index, where the first mutual information is one of the at least one piece of third mutual information, the second mutual information is one of the at least one piece of fourth mutual information, the first ratio is one of the at least one second ratio, the first neuron node size is one of the at least one second neuron node size, and the first ratio range is one of the at least one second ratio range.

[0039] In a possible implementation, the first mutual information is an amount of information about an input piece of in an output of the first AI model, the second mutual information is an amount of information about an output piece of in an input of the second AI model, the first ratio is a ratio of the second mutual information to the first mutual information, the first neuron node size is configured to indicate an output format of the first AI model or an input format of the second AI model, and the first ratio range is a range of multiple first ratios.

[0040] In a possible implementation, the receiving first information includes: receiving a model switch request.

[0041] In a possible implementation, the method further includes: sending a second message indicating a method for calculating the first information.

[0042] In a possible implementation, the method for calculating the first information is Hilbert-Schmidt independence criterion (HSIC), or a predefined mutual information approximation method.

[0043] In a possible implementation, the sending a first message includes: sending the first message when a value of the first information is not within a corresponding predetermined range.

[0044] In a possible implementation, if the ratio or ratio range or mutual information of the UE is out of a corresponding predetermined range, the UE can send a model switch request to the BS. The BS receives the model switch request and sends the first message indicating the UE to perform the switching of the model.

[0045] In a possible implementation, the method further includes adjusting the AI model at a network device side based on the first information.

[0046] The BS can adjust its own AI model to adapt to the AI model of the UE based on the first information sent by the UE.

[0047] In a possible implementation, the first AI model is a decoder and the second AI model is an encoder; or, the first AI model is an encoder and the second AI model is a decoder.

[0048] In a possible implementation, the first AI model is at a user equipment side and the second AI model is at a network device side; or the first AI model and the second AI model are at a user equipment side.

[0049] In a possible implementation, the method is executed by a network device.

[0050] For the beneficial effects of the second aspect, reference is made to the first aspect. Details are not described

herein again.

[0051] According to a third aspect, this application provides a communication apparatus, including: a sending module configured to send first information related to mutual information of a first artificial intelligence (AI) model and a second AI model, the first AI model and the second AI model constituting a two-sided model; and a receiving module configured to receive a first message indicating an AI model related to the first AI model.

[0052] In a possible implementation, the first information includes at least one of first mutual information, second mutual information, a first ratio, a first neuron node size, and a first ratio range.

[0053] In a possible implementation, the first information includes an index corresponding to first mutual information, an index corresponding to second mutual information, an index corresponding to a first ratio, an index corresponding to a first neuron node size, or an index corresponding to a first ratio range.

[0054] In a possible implementation, at least one piece of third mutual information, at least one piece of fourth mutual information, at least one second ratio, at least one second neuron node size or at least one second ratio range is predetermined or configured by a network device, where each of third mutual information, fourth mutual information, second ratios, second neuron node sizes or second ratio ranges corresponds to an index, where the first mutual information is one of the at least one piece of third mutual information, the second mutual information is one of the at least one piece of fourth mutual information, the first ratio is one of the at least one second ratio, the first neuron node size is one of the at least one second neuron node size, and the first ratio range is one of the at least one second ratio range.

[0055] In a possible implementation, the first mutual information is an amount of information about an input included in an output of the first AI model, the second mutual information is an amount of information about an output included in an input of the second AI model, the first ratio is a ratio of the second mutual information to the first mutual information, the first neuron node size is configured to indicate an output format of the first AI model or an input format of the second AI model, and the first ratio range is a range of multiple first ratios.

[0056] In a possible implementation, the sending module is further configured to send the first information when the first information has changed during a time period.

[0057] In a possible implementation, the sending module is further configured to send a model switch request.

[0058] In a possible implementation, the receiving module is further configured to receive a second message indicating a method for calculating the first information.

[0059] In a possible implementation, the method for calculating the first information is Hilbert-Schmidt independence criterion (HSIC), or a predefined mutual information approximation method.

[0060] In a possible implementation, the first AI model is a decoder and the second AI model is an encoder; or, the first

AI model is an encoder and the second AI model is a decoder.

[0061] In a possible implementation, the first AI model is at a user equipment side and the second AI model is at a network device side; or the first AI model and the second AI model are at a user equipment side.

[0062] In a possible implementation, the apparatus is located on a user equipment.

[0063] According to a fourth aspect, this application provides a communication apparatus, including: a receiving module configured to receive first information related to mutual information of a first artificial intelligence (AI) model and a second AI model, the first AI model and the second AI model constituting a two-sided model; and a sending module configured to send a first message indicating an AI model related to the first AI model.

[0064] In a possible implementation, the first information includes at least one of first mutual information, second mutual information, a first ratio, a first neuron node size, and a first ratio range.

[0065] In a possible implementation, the first information includes an index corresponding to first mutual information, an index corresponding to second mutual information, an index corresponding to a first ratio, an index corresponding to a first neuron node size, or an index corresponding to a first ratio range.

[0066] In a possible implementation, at least one piece of third mutual information, at least one piece of fourth mutual information, at least one second ratio, at least one second neuron node size or at least one second ratio range is predetermined or configured by a network device, where each of third mutual information, fourth mutual information, second ratios, second neuron node sizes or second ratio ranges corresponds to an index, where the first mutual information is one of the at least one piece of third mutual information, the second mutual information is one of the at least one piece of fourth mutual information, the first ratio is one of the at least one second ratio, the first neuron node size is one of the at least one second neuron node size, and the first ratio range is one of the at least one second ratio range.

[0067] In a possible implementation, the first mutual information is an amount of information about an input piece of in an output of the first AI model, the second mutual information is an amount of information about an output piece of in an input of the second AI model, the first ratio is a ratio of the second mutual information to the first mutual information, the first neuron node size is configured to indicate an output format of the first AI model or an input format of the second AI model, and the first ratio range is a range of multiple first ratios.

[0068] In a possible implementation, the sending module is further configured to send the first information when the first information has changed during a time period.

[0069] In a possible implementation, the sending module is further configured to send a model switch request.

[0070] In a possible implementation, the receiving module is further configured to receive a second message indicating a method for calculating the first information.

[0071] In a possible implementation, the method for calculating the first information is Hilbert-Schmidt independence criterion (HSIC), or a predefined mutual information approximation method.

[0072] In a possible implementation, the first AI model is a decoder and the second AI model is an encoder; or, the first AI model is an encoder and the second AI model is a decoder.

[0073] In a possible implementation, the first AI model is at a user equipment side and the second AI model is at a network device side; or the first AI model and the second AI model are at a user equipment side.

[0074] In a possible implementation, the apparatus is located on a user equipment.

[0075] According to a fifth aspect, a communication apparatus including a processor and a memory is provided. The processor is connected to the memory. The memory is configured to store instructions, and the processor is configured to execute the instructions. When the processor executes the instructions stored in the memory, the processor is enabled to perform the method in any possible implementation of the first aspect or the second aspect.

[0076] According to a sixth aspect, this application provides a communication system, which includes the communication apparatus in any possible implementation of the third aspect, as well as the communication apparatus in any possible implementation of the fourth aspect.

[0077] According to a seventh aspect, this application provides a computer readable storage medium, which includes instructions. When the instructions run on a processor, the processor is enabled to perform the method in any possible implementation of the first aspect or the second aspect.

[0078] According to an eighth aspect, this application provides a computer program product, which includes computer program code. When the computer program code runs on a computer, the computer is enabled to perform the method in any possible implementation of the first aspect or the second aspect.

[0079] It should be noted that all or a part of the above computer program code can be stored in on a first storage medium. The first storage medium can be packaged together with the processor or separately with the processor.

[0080] According to a ninth aspect, this application provides a chip system, which includes a memory and a processor. The memory is configured to store a computer program, and the processor is configured to invoke the computer program from the memory and run the computer program, so that an electronic device on which the chip system is disposed performs the method in any possible implementation of the first aspect or the second aspect.

DESCRIPTION OF DRAWINGS

[0081] FIG. 1 is a schematic diagram of a communication system according to an embodiment of the present application.

[0082] FIG. 2 is a schematic diagram of a communication system 100 according to an embodiment of the present application.

[0083] FIG. 3 is a schematic diagram of an ED 110 and a base station 170a, 170b and/or 170c according to an embodiment of the present application.

[0084] FIG. 4 is a schematic diagram of units or modules in a device according to an embodiment of the present application.

[0085] FIG. 5 is a schematic diagram of an AI-based communication device.

[0086] FIG. 6 is a schematic diagram of a device 500 receiving reference data samples from a device 600 according to an embodiment of the present application.

[0087] FIG. 7 is a schematic diagram of reference data samples consisting of a plurality of groups according to an embodiment of the present application.

[0088] FIG. 8 is a schematic representation of a DNN-based approximation according to an embodiment of the present application.

[0089] FIG. 9 is a schematic diagram of an information bottleneck as a learning ratio according to an embodiment of the present application.

[0090] FIG. 10 is a schematic diagram of autoencoders including the cross check on the learning metric ratios according to an embodiment of the present application.

[0091] FIG. 11 is a schematic diagram of a split encoder and decoder according to an embodiment of the present application.

[0092] FIG. 12 is a schematic diagram of an auto-encoder model according to an embodiment of the present application.

[0093] FIG. 13 is a flowchart of a communication method according to an embodiment of the present application.

[0094] FIG. 14 is a flowchart of a communication method according to an embodiment of the present application.

[0095] FIG. 15 is a schematic diagram of the UE calculating mutual information or mutual information ratios according to an embodiment of the present application.

[0096] FIG. 16 is a flowchart of a communication method according to an embodiment of the present application.

[0097] FIG. 17 is a flowchart of a communication method according to an embodiment of the present application.

[0098] FIG. 18 is a schematic block diagram of a communication apparatus 1800 according to an embodiment of this application.

[0099] FIG. 19 is a schematic block diagram of a communication apparatus 1900 according to an embodiment of this

application.

[00100] FIG. 20 is a schematic block diagram of still another communication apparatus according to an embodiment of the present application.

DESCRIPTION OF EMBODIMENTS

[00101] The following describes the technical solutions in the present application with reference to the accompanying drawings.

[00102] The following describes the technical solutions in the present application with reference to the accompanying drawings. Obviously, the described embodiments are part of the embodiments of the present application, and not all of them. Based on the embodiments in the present application, all other embodiments obtained by a person of ordinary skill in the art without making creative labor shall fall within the scope of protection of the present application.

[00103] The present application will present aspects, embodiments, or features around systems that include multiple devices, components, modules, etc. It should be understood and appreciated that the individual systems may include additional devices, components, modules, etc., and/or may not include all of the devices, components, modules, etc, discussed in connection with the accompanying drawings. In addition, combinations of these options may be used.

[00104] In addition, in the embodiments of the present application, the word "exemplarily" and the phrase "as an example" are used to indicate, for example, illustration or description. Any embodiment or design solution described as "exemplarily" in this application should not be construed as being superior to or more advantageous than other embodiments or design solutions. Rather, the use of the word "example" is intended to present the concept in a specific manner.

[00105] The phrases "in some possible embodiments", "in some possible application scenarios", etc., appearing in various places in this description, do not necessarily refer to the same embodiments, but rather mean "one or more, but not all, embodiments" unless otherwise specifically emphasized. Unless otherwise specifically emphasized, the terms "including", "comprising", "having", and variations thereof all mean "including but not limited to".

[00106] In the present application, "at least one" refers to one or more, and "multiple" refers to two or more. "and/or", describing the association of the associated objects, indicates that three relationships can exist. For example, A and/or B can mean A alone, both A and B, and B alone, where A and B can be singular or plural. The character "/" generally indicates that the preceding and following associated objects are in an "or" relationship.

[00107] The application scenarios described in the embodiments of the present application are intended to illustrate the technical solutions of the embodiments of the present application more clearly and do not constitute a limitation to the technical

solutions provided by the embodiments of the present application. It is known to those of ordinary skill in the art that the technical solutions provided by the embodiments of the present application are equally applicable to similar technical problems as the system architecture evolves and new application scenarios emerge.

[00108] The technical solutions in embodiments of this application may be applied to various communications systems, such as a Global System for Mobile Communications (GSM), a Code Division Multiple Access (CDMA) system, a Wideband Code Division Multiple Access (WCDMA) system, a general packet radio service (GPRS) system, a Long Term Evolution (LTE) system, an LTE frequency division duplex (FDD) system, an LTE time division duplex (TDD) system, a Universal Mobile Telecommunications System (UMTS), a Worldwide Interoperability for Microwave Access (WiMAX) communications system, a wireless local area network (WLAN), a fifth generation (5G) wireless communications system, a new ratio (NR) wireless communications system, a sixth generation (6G) wireless communications system, or other evolving communications systems.

[00109] In order to better describe the solutions of embodiments in the present application, concepts and terms that may be involved in the present application will be described below.

[00110] (1) Data collection

[00111] Data is a very important component for artificial intelligence (AI)/machine learning (ML) techniques. Data collection is a process of collecting data by the network nodes, management entity, or UE for the purpose of AI/ML model training, data analytics and inference.

[00112] (2) AI/ML model training

[00113] AI/ML model training is a process to train an AI/ML model by learning the input/output relationship in a data driven manner and obtain the trained AI/ML Model for inference.

[00114] (3) AI/ML model inference

[00115] A process of using a trained AI/ML model to produce a set of outputs based on a set of inputs.

[00116] (4) AI/ML model validation

[00117] As a sub-process of training, validation is used to evaluate the quality of an AI/ML model using a dataset different from the one used for model training. Validation can help select model parameters that generalize beyond the dataset used for model training. The model parameter after training can be adjusted further by the validation process.

[00118] (5) AI/ML model testing

[00119] Similar to validation, testing is also a sub-process of training, and it is used to evaluate the performance of a final AI/ML model using a dataset different from the one used for model training and validation. Different from AI/ML model validation, testing does not assume subsequent tuning of the model.

[00120] (6) Online training

[00121] Online training means an AI/ML training process where the model being used for inference is typically continuously trained in (near) real-time with the arrival of new training samples.

[00122] (7) Offline training

[00123] Offline training is an AI/ML training process where the model is trained based on the collected dataset, and where the trained model is later used or delivered for inference.

[00124] (8) AI/ML model delivery/transfer

[00125] AI/ML model delivery/transfer is a generic term referring to the delivery of an AI/ML model from one entity to another entity in any manner. Delivery of an AI/ML model over the air interface includes either parameters of a model structure known at the receiving end or a new model with parameters. Delivery may contain a full model or a partial model.

[00126] (9) Life cycle management (LCM)

[00127] When the AI/ML model is trained and/or inferred at one device, it is necessary to monitor and manage the whole AI/ML process to guarantee the performance gain obtained by AI/ML technologies. For example, due to the randomness of wireless channels and the mobility of UEs, the propagation environment of wireless signals changes frequently. Nevertheless, it is difficult for an AI/ML model to maintain optimal performance in all scenarios for all the time, and the performance may even deteriorate sharply in some scenarios. Therefore, the lifecycle management (LCM) of AI/ML models is essential for sustainable operation of AI/ML in the NR air-interface. Life cycle management covers the whole procedure of AI/ML technologies applied on one or more nodes. In specific, it includes at least one of the following sub-process: data collection, model training, model identification, model registration, model deployment, model configuration, model inference, model selection, model activation, deactivation, model switching, model fallback, model monitoring, model update, model transfer/delivery, and UE capability report. Model monitoring can be based on inference accuracy, including metrics related to intermediate key performance indicators (KPIs), and it can also be based on system performance, including metrics related to system performance KPIs, e.g., accuracy and relevance, overhead, complexity (computation and memory cost), latency (timeliness of monitoring result, from model failure to action) and power consumption. Moreover, data distribution may shift after deployment due to environmental changes, and thus the model based on input or output data distribution should also be considered.

[00128] (10) Supervised learning

[00129] The goal of supervised learning algorithms is to train a model that maps feature vectors (inputs) to labels (output), based on the training data which includes the example feature-label pairs. The supervised learning can analyze the training data and produce an inferred function, which can be used for mapping the inference data. Supervised learning can be further divided

into two types: Classification and Regression. Classification is used when the output of the AI/ML model is categorical i.e., with two or more classes. Regression is used when the output of the AI/ML model is a real or continuous value.

[00130] (11) Unsupervised learning

[00131] In contrast to supervised learning where the AI/ML models learn to map the input to the target output, the unsupervised methods learn concise representations of the input data without the labelled data, which can be used for data exploration or to analyze or generate new data. One typical unsupervised learning is clustering which explores the hidden structure of input data and provides the classification results for the data.

[00132] (12) Reinforcement learning

[00133] Reinforcement learning is used to solve sequential decision-making problems. Reinforcement learning is a process of training the action of an intelligent agent from input (state) and a feedback signal (reward) in an environment. In reinforcement learning, an intelligent agent interacts with an environment by taking an action to maximize the cumulative reward. Whenever the intelligent agent takes one action, the current state in the environment may transfer to the new state, and the new state resulting from the action will bring the associated reward. Then the intelligent agent can take the next action based on the received reward and new state in the environment. During the training phase, the agent interacts with the environment to collect experience. The environments are often mimicked by the simulator since it is expensive to directly interact with the real system. In the inference phase, the agent can use the optimal decision-making rule learned from the training phase to achieve the maximal accumulated reward.

[00134] (13) Federated learning

[00135] Federated learning (FL) is a machine learning technique that is used to train an AI/ML model by a central node (e.g., server) and a plurality of decentralized edge nodes (e.g., UEs, next Generation NodeBs, "gNBs"). According to the wireless FL technique, a server may provide, to an edge node, a set of model parameters (e.g., weights, biases, gradients) that describe a global AI/ML model. The edge node may initialize a local AI/ML model with the received global AI/ML model parameters. The edge node may then train the local AI/ML model using local data samples to, thereby, produce a trained local AI/ML model. The edge node may then provide, to the serve, a set of AI/ML model parameters that describe the local AI/ML model. Upon receiving, from a plurality of edge nodes, a plurality of sets of AI/ML model parameters that describe respective local AI/ML models at the plurality of edge nodes, the server may aggregate the local AI/ML model parameters reported from the plurality of UEs and, based on such aggregation, update the global AI/ML model. A subsequent iteration progresses much like the first iteration. The server may transmit the aggregated global model to a plurality of edge nodes. The above procedure is performed multiple iterations until the global AI/ML model is considered to be finalized, e.g., the AI/ML model is converged or the training stopping conditions are satisfied. Notably, the wireless FL technique does not involve the exchange of local data

samples. Indeed, the local data samples remain at respective edge nodes.

[00136] AI-based algorithms have been introduced into modern wireless communications to solve some wireless problems such as channel estimation, scheduling, channel state information (CSI) compression (from user equipment to base station), Multiple-in Multiple-Out (MIMO)'s beamforming, positioning, and so on. AI algorithm is a data-driven method that tunes some predefined architectures by a set of data samples called as training data set. The recent AI trains deep neural network (DNN) (including CNN, RNN, transformer, etc.) architecture by setting the neurons with a SGD algorithm.

[00137] AI techniques (including ML techniques) in communication include AI-based communications in the physical layer and/or AI-based communications in the MAC layer. For the physical layer, the AI communication may aim to optimize component design and/or improve algorithm performance. For the MAC layer, the AI/ML based communication may aim to utilize the AI/ML capability for learning, prediction, and/or making a decision to solve a complicated optimization problem with possible better strategy and/or optimal solution, e.g. to optimize the functionality in the MAC layer, e.g. intelligent TRP management, intelligent beam management, intelligent channel resource allocation, intelligent power control, intelligent spectrum utilization, intelligent modulation and coding scheme (MCS), intelligent hybrid automatic repeat request (HARQ) strategy, intelligent transmit/receive (Tx/Rx) mode adaption, etc.

[00138] AI architecture may involve multiple nodes, where the multiple nodes may be organized in one of two modes, i.e., centralized and distributed, both of which may be deployed in an access network, a core network, or an edge computing system, or a third party network. A centralized training and computing architecture is restricted by possibly large communication overhead and strict user data privacy. A distributed training and computing architecture may include several frameworks, e.g., distributed machine learning and federated learning. In some embodiments, an AI architecture may include an intelligent controller which can perform as a single agent or a multi-agent, based on joint optimization or individual optimization. New protocols and signaling mechanisms are desired so that the corresponding interface link can be personalized with customized parameters to meet particular requirements while minimizing signaling overhead and maximizing the whole system spectrum efficiency by personalized AI technologies.

[00139] New protocols and signaling mechanisms are provided for operating within and switching between different modes of operation, including between AI and non-AI modes, and for measurement and feedback to accommodate the different possible measurements and information that may need to be fed back, depending upon the implementation.

[00140] It is now quite common for neural network models to become larger and deeper, which may easily require more computational resources than just one or two computers. Most neural network models would be trained on a powerful computation cloud. A user with a desired neural network architecture, raw training data set, and training goal may not have sufficient local computation resources to train their model locally. In order to access a powerful computation cloud, the user

would have to transmit all the specifications of its neural network architecture, its training data set, and its training goal to the network cloud completely. It is mandated that the user must trust the cloud and grant the cloud full authorization to manipulate its intellectual property (neural network architecture, training data set, and training goal).

[00141] As data-driven method, AI-based algorithms inevitably suffer from low generalization: if a testing data sample were an outlier to the training data set, a neural network wouldn't make a good inference on the test data sample. Even if the AI model is trained on a large number of data sets, it may also not possess the necessary knowledge to perform effectively in other environments, especially in wireless communication where the channel information is changed rapidly.

[00142] In the present application, the AI model is exemplified by a DNN, i.e., a deep neural network or network. The specific AI model should not be construed as a limitation of the present application.

[00143] FIG. 1 is a schematic diagram of a communication system according to an embodiment of the present application.

[00144] Referring to FIG.1, as an illustrative example without limitation, a simplified schematic illustration of a communication system is provided. The communication system 100 includes a radio access network 120. The radio access network 120 may be a next generation (e.g. sixth generation (6G) or later) radio access network, or a legacy (e.g. 5G, 4G, 3G or 2G) radio access network. One or more communication electric devices (EDs) 110a-120j (generically referred to as 110) may be interconnected to one another or connected to one or more network nodes (170a, 170b, generically referred to as 170) in the radio access network 120. A core network 130 may be a part of the communication system and may be dependent or independent of the radio access technology used in the communication system 100. Also, the communication system 100 includes a public switched telephone network (PSTN) 140, the internet 150, and other networks 160.

[00145] FIG. 2 is a schematic diagram of a communication system 100 according to an embodiment of the present application.

[00146] FIG. 2 illustrates an example communication system 100. In general, the communication system 100 enables multiple wireless or wired elements to communicate data and other content. The purpose of the communication system 100 may be to provide content, such as voice, data, video, and/or text, via broadcast, multicast and unicast, etc. The communication system 100 may operate by sharing resources, such as carrier spectrum bandwidth, between its constituent elements. The communication system 100 may include a terrestrial communication system and/or a non-terrestrial communication system. The communication system 100 may provide a wide range of communication services and applications (such as earth monitoring, remote sensing, passive sensing and positioning, navigation and tracking, autonomous delivery and mobility, etc.). The communication system 100 may provide a high degree of availability and robustness through a joint operation of the terrestrial communication system and the non-terrestrial communication system. For example, integrating a non-terrestrial communication system (or components thereof) into a terrestrial communication system can result in what may be considered

a heterogeneous network including multiple layers. Compared to conventional communication networks, the heterogeneous network may achieve better overall performance through efficient multi-link joint operation, more flexible functionality sharing, and faster physical layer link switching between terrestrial networks and non-terrestrial networks.

[00147] The terrestrial communication system and the non-terrestrial communication system can be regarded as sub-systems of the communication system. In the example shown, the communication system 100 includes electronic devices (EDs) 110a-110d (generically referred to as ED 110), radio access networks (RANs) 120a-120b, non-terrestrial communication network 120c, a core network 130, a public switched telephone network (PSTN) 140, the internet 150, and other networks 160. The RANs 120a-120b include respective base stations (BSs) 170a-170b, which may be generically referred to as terrestrial transmit and receive points (T-TRPs) 170a-170b. The non-terrestrial communication network 120c includes an access node 120c, which may be generically referred to as a non-terrestrial transmit and receive point (NT-TRP) 172.

[00148] Any ED 110 may be alternatively or additionally configured to interface, access, or communicate with any other T-TRP 170a-170b and NT-TRP 172, the internet 150, the core network 130, the PSTN 140, the other networks 160, or any combination of the preceding. In some examples, ED 110a may communicate an uplink and/or downlink transmission over an interface 190a with T-TRP 170a. In some examples, the EDs 110a, 110b and 110d may also communicate directly with one another via one or more sidelink air interfaces 190b. In some examples, ED 110d may communicate an uplink and/or downlink transmission over an interface 190c with NT-TRP 172.

[00149] The air interfaces 190a and 190b may use similar communication technology, such as any suitable radio access technology. For example, the communication system 100 may implement one or more channel access methods, such as code division multiple access (CDMA), time division multiple access (TDMA), frequency division multiple access (FDMA), orthogonal FDMA (OFDMA), or single-carrier FDMA (SC-FDMA) in the air interfaces 190a and 190b. The air interfaces 190a and 190b may utilize other higher dimension signal spaces, which may involve a combination of orthogonal and/or non-orthogonal dimensions. The air interface 190c can enable communication between the ED 110d and one or multiple NT-TRPs 172 via a wireless link or simply a link. For some examples, the link is a dedicated connection for unicast transmission, a connection for broadcast transmission, or a connection between a group of EDs and one or multiple NT-TRPs for multicast transmission.

[00150] The RANs 120a and 120b are in communication with the core network 130 to provide the EDs 110a 110b, and 110c with various services such as voice, data, and other services. The RANs 120a and 120b and/or the core network 130 may be in direct or indirect communication with one or more other RANs (not shown), which may or may not be directly served by core network 130, and may or may not employ the same radio access technology as RAN 120a, RAN 120b or both. The core network 130 may also serve as a gateway access between (i) the RANs 120a and 120b or EDs 110a 110b, and 110c or both,

and (ii) other networks (such as the PSTN 140, the internet 150, and the other networks 160). In addition, some or all of the EDs 110a 110b, and 110c may include functionality for communicating with different wireless networks over different wireless links using different wireless technologies and/or protocols. Instead of wireless communication (or in addition thereto), the EDs 110a 110b, and 110c may communicate via wired communication channels to a service provider or switch (not shown), and to the internet 150. PSTN 140 may include circuit switched telephone networks for providing plain old telephone service (POTS). Internet 150 may include a network of computers and subnets (intranets) or both, and incorporate protocols, such as internet protocol (IP), transmission control protocol (TCP), and user datagram protocol (UDP). EDs 110a 110b, and 110c may be multimode devices capable of operation according to multiple radio access technologies, and incorporate multiple transceivers necessary to support such.

[00151] FIG. 3 is a schematic diagram of an ED 110 and a base station 170a, 170b and/or 170c according to an embodiment of the present application.

[00152] FIG. 3 illustrates another example of an ED 110 and a base station 170a, 170b and/or 170c. The ED 110 is used to connect persons, objects, machines, etc. The ED 110 may be widely used in various scenarios, for example, cellular communications, device-to-device (D2D), vehicle to everything (V2X), peer-to-peer (P2P), machine-to-machine (M2M), machine-type communications (MTC), internet of things (IoT), virtual reality (VR), augmented reality (AR), industrial control, self-driving, remote medical, smart grid, smart furniture, smart office, smart wearable, smart transportation, smart city, drones, robots, remote sensing, passive sensing, positioning, navigation and tracking, autonomous delivery and mobility, etc.

[00153] Each ED 110 represents any suitable end user device for wireless operation and may include such devices (or may be referred to) as a user equipment/device (UE), a wireless transmit/receive unit (WTRU), a mobile station, a fixed or mobile subscriber unit, a cellular telephone, a station (STA), a machine type communication (MTC) device, a personal digital assistant (PDA), a smartphone, a laptop, a computer, a tablet, a wireless sensor, a consumer electronics device, a smart book, a vehicle, a car, a truck, a bus, a train, or an IoT device, an industrial device, or apparatus (e.g. communication module, modem, or chip) in the forgoing devices, among other possibilities. Future generation EDs 110 may be referred to using other terms. The base station 170a and 170b is a T-TRP and will hereafter be referred to as T-TRP 170. Also shown in FIG.3, a NT-TRP will hereafter be referred to as NT-TRP 172. Each ED 110 connected to T-TRP 170 and/or NT-TRP 172 can be dynamically or semi-statically turned on (i.e., established, activated, or enabled), turned off (i.e., released, deactivated, or disabled) and/or configured in response to one or more of connection availability and connection necessity.

[00154] The ED 110 includes a transmitter 201 and a receiver 203 coupled to one or more antennas 204. Only one antenna 204 is illustrated. One, some, or all of the antennas may alternatively be panels. The transmitter 201 and the receiver 203 may be integrated, e.g. as a transceiver. The transceiver is configured to modulate data or other content for transmission

by at least one antenna 204 or network interface controller (NIC). The transceiver is also configured to demodulate data or other content received by the at least one antenna 204. Each transceiver includes any suitable structure for generating signals for wireless or wired transmission and/or processing signals received wirelessly or by wire. Each antenna 204 includes any suitable structure for transmitting and/or receiving wireless or wired signals.

[00155] The ED 110 includes at least one memory 208. The memory 208 stores instructions and data used, generated, or collected by the ED 110. For example, the memory 208 can store software instructions or modules configured to implement some or all of the functionality and/or embodiments described herein and that are executed by the processing unit(s) 210. Each memory 208 includes any suitable volatile and/or non-volatile storage and retrieval device(s). Any suitable type of memory may be used, such as random access memory (RAM), read only memory (ROM), hard disk, optical disc, subscriber identity module (SIM) card, memory stick, secure digital (SD) memory card, on-processor cache, and the like.

[00156] The ED 110 may further include one or more input/output devices (not shown) or interfaces (such as a wired interface to the internet 150 in FIG. 1). The input/output devices permit interaction with a user or other devices in the network. Each input/output device includes any suitable structure for providing information to or receiving information from a user, such as a speaker, microphone, keypad, keyboard, display, or touch screen, including network interface communications.

[00157] The ED 110 further includes a processor 210 for performing operations including those related to preparing a transmission for uplink transmission to the NT-TRP 172 and/or T-TRP 170, those related to processing downlink transmissions received from the NT-TRP 172 and/or T-TRP 170, and those related to processing sidelink transmission to and from another ED 110. Processing operations related to preparing a transmission for uplink transmission may include operations such as encoding, modulating, transmit beamforming, and generating symbols for transmission. Processing operations related to processing downlink transmissions may include operations such as receive beamforming, demodulating and decoding received symbols. Depending upon the embodiment, a downlink transmission may be received by the receiver 203, possibly using receive beamforming, and the processor 210 may extract signaling from the downlink transmission (e.g. by detecting and/or decoding the signaling). An example of signaling may be a reference signal transmitted by NT-TRP 172 and/or T-TRP 170. In some embodiments, the processor 276 implements the transmit beamforming and/or receive beamforming based on the indication of beam direction, e.g. beam angle information (BAI), received from T-TRP 170. In some embodiments, the processor 210 may perform operations relating to network access (e.g. initial access) and/or downlink synchronization, such as operations relating to detecting a synchronization sequence, decoding and obtaining the system information, etc. In some embodiments, the processor 210 may perform channel estimation, e.g. using a reference signal received from the NT-TRP 172 and/or T-TRP 170.

[00158] Although not illustrated, the processor 210 may form part of the transmitter 201 and/or receiver 203. Although

not illustrated, the memory 208 may form part of the processor 210.

[00159] The processor 210, and the processing components of the transmitter 201 and receiver 203 may each be implemented by the same or different one or more processors that are configured to execute instructions stored in a memory (e.g. in memory 208). Alternatively, some or all of the processor 210, and the processing components of the transmitter 201 and receiver 203 may be implemented using dedicated circuitry, such as a programmed field-programmable gate array (FPGA), a graphical processing unit (GPU), or an application-specific integrated circuit (ASIC).

[00160] The T-TRP 170 may be known by other names in some implementations, such as a base station, a base transceiver station (BTS), a radio base station, a network node, a network device, a device on the network side, a transmit/receive node, a Node B, an evolved NodeB (eNodeB or eNB), a Home eNodeB, a next Generation NodeB (gNB), a transmission point (TP), a site controller, an access point (AP), or a wireless router, a relay station, a remote radio head, a terrestrial node, a terrestrial network device, or a terrestrial base station, base band unit (BBU), remote radio unit (RRU), active antenna unit (AAU), remote radio head (RRH), central unit (CU), distribute unit (DU), positioning node, among other possibilities. The T-TRP 170 may be macro BSs, pico BSs, relay nodes, donor nodes, or the like, or combinations thereof. The T-TRP 170 may refer to the forging devices or apparatus (e.g. communication module, modem, or chip) in the forgoing devices.

[00161] In some embodiments, the parts of the T-TRP 170 may be distributed. For example, some of the modules of the T-TRP 170 may be located remote from the equipment housing the antennas of the T-TRP 170, and may be coupled to the equipment housing the antennas over a communication link (not shown) sometimes known as front haul, such as common public radio interface (CPRI). Therefore, in some embodiments, the term T-TRP 170 may also refer to modules on the network side that perform processing operations, such as determining the location of the ED 110, resource allocation (scheduling), message generation, and encoding/decoding, and that are not necessarily part of the equipment housing the antennas of the T-TRP 170. The modules may also be coupled to other T-TRPs. In some embodiments, the T-TRP 170 may actually be a plurality of T-TRPs that are operating together to serve the ED 110, e.g. through coordinated multipoint transmissions.

[00162] The T-TRP 170 includes at least one transmitter 252 and at least one receiver 254 coupled to one or more antennas 256. Only one antenna 256 is illustrated. One, some, or all of the antennas may alternatively be panels. The transmitter 252 and the receiver 254 may be integrated as a transceiver. The T-TRP 170 further includes a processor 260 for performing operations including those related to: preparing a transmission for downlink transmission to the ED 110, processing an uplink transmission received from the ED 110, preparing a transmission for backhaul transmission to NT-TRP 172, and processing a transmission received over backhaul from the NT-TRP 172. Processing operations related to preparing a transmission for downlink or backhaul transmission may include operations such as encoding, modulating, precoding (e.g. MIMO precoding), transmit beamforming, and generating symbols for transmission. Processing operations related to processing received

transmissions in the uplink or over backhaul may include operations such as receive beamforming, and demodulating and decoding received symbols. The processor 260 may also perform operations relating to network access (e.g. initial access) and/or downlink synchronization, such as generating the content of synchronization signal blocks (SSBs), generating the system information, etc. In some embodiments, the processor 260 also generates the indication of beam direction, e.g. BAI, which may be scheduled for transmission by scheduler 253. The processor 260 performs other network-side processing operations described herein, such as determining the location of the ED 110, determining where to deploy NT-TRP 172, etc. In some embodiments, the processor 260 may generate signaling, e.g. to configure one or more parameters of the ED 110 and/or one or more parameters of the NT-TRP 172. Any signaling generated by the processor 260 is sent by the transmitter 252. Note that "signaling", as used herein, may alternatively be called control signaling. Dynamic signaling may be transmitted in a control channel, e.g. a physical downlink control channel (PDCCH), and static or semi-static higher layer signaling may be included in a packet transmitted in a data channel, e.g. in a physical downlink shared channel (PDSCH).

[00163] A scheduler 253 may be coupled to the processor 260. The scheduler 253 may be included within or operated separately from the T-TRP 170, which may schedule uplink, downlink, and/or backhaul transmissions, including issuing scheduling grants and/or configuring scheduling-free ("configured grant") resources. The T-TRP 170 further includes a memory 258 for storing information and data. The memory 258 stores instructions and data used, generated, or collected by the T-TRP 170. For example, the memory 258 can store software instructions or modules configured to implement some or all of the functionality and/or embodiments described herein and that are executed by the processor 260.

[00164] Although not illustrated, the processor 260 may form part of the transmitter 252 and/or receiver 254. Also, although not illustrated, the processor 260 may implement the scheduler 253. Although not illustrated, the memory 258 may form part of the processor 260.

[00165] The processor 260, the scheduler 253, and the processing components of the transmitter 252 and receiver 254 may each be implemented by the same or different one or more processors that are configured to execute instructions stored in a memory, e.g. in memory 258. Alternatively, some or all of the processor 260, the scheduler 253, and the processing components of the transmitter 252 and receiver 254 may be implemented using dedicated circuitry, such as a FPGA, a GPU, or an ASIC.

[00166] Although the NT-TRP 172 is illustrated as a drone only as an example, the NT-TRP 172 may be implemented in any suitable non-terrestrial form. Also, the NT-TRP 172 may be known by other names in some implementations, such as a non-terrestrial node, a non-terrestrial network device, or a non-terrestrial base station. The NT-TRP 172 includes a transmitter 272 and a receiver 274 coupled to one or more antennas 280. Only one antenna 280 is illustrated. One, some, or all of the antennas may alternatively be panels. The transmitter 272 and the receiver 274 may be integrated as a transceiver. The NT-TRP

172 further includes a processor 276 for performing operations including those related to: preparing a transmission for downlink transmission to the ED 110, processing an uplink transmission received from the ED 110, preparing a transmission for backhaul transmission to T-TRP 170, and processing a transmission received over backhaul from the T-TRP 170. Processing operations related to preparing a transmission for downlink or backhaul transmission may include operations such as encoding, modulating, precoding (e.g. MIMO precoding), transmit beamforming, and generating symbols for transmission. Processing operations related to processing received transmissions in the uplink or over backhaul may include operations such as receive beamforming, and demodulating and decoding received symbols. In some embodiments, the processor 276 implements the transmit beamforming and/or receive beamforming based on beam direction information (e.g. BAI) received from T-TRP 170. In some embodiments, the processor 276 may generate signaling, e.g. to configure one or more parameters of the ED 110. In some embodiments, the NT-TRP 172 implements physical layer processing, but does not implement higher layer functions such as functions at the medium access control (MAC) or radio link control (RLC) layer. As this is only an example, more generally, the NT-TRP 172 may implement higher layer functions in addition to physical layer processing.

[00167] The NT-TRP 172 further includes a memory 278 for storing information and data. Although not illustrated, the processor 276 may form part of the transmitter 272 and/or receiver 274. Although not illustrated, the memory 278 may form part of the processor 276.

[00168] The processor 276 and the processing components of the transmitter 272 and receiver 274 may each be implemented by the same or different one or more processors that are configured to execute instructions stored in a memory, e.g. in memory 278. Alternatively, some or all of the processor 276 and the processing components of the transmitter 272 and receiver 274 may be implemented using dedicated circuitry, such as a programmed FPGA, a GPU, or an ASIC. In some embodiments, the NT-TRP 172 may actually be a plurality of NT-TRPs that are operating together to serve the ED 110, e.g. through coordinated multipoint transmissions.

[00169] The T-TRP 170, the NT-TRP 172, and/or the ED 110 may include other components, but these have been omitted for the sake of clarity.

[00170] FIG. 4 is a schematic diagram of units or modules in a device according to an embodiment of the present application.

[00171] One or more steps of the embodiment methods provided may be performed by corresponding units or modules, according to FIG. 4. FIG. 4 illustrates units or modules in a device, such as in ED 110, T-TRP 170, or NT-TRP 172. For example, a signal may be transmitted by a transmitting unit or a transmitting module. For example, a signal may be transmitted by a transmitting unit or a transmitting module. A signal may be received by a receiving unit or a receiving module. A signal may be processed by a processing unit or a processing module. Other steps may be performed by an artificial intelligence (AI) or

machine learning (ML) module. The respective units or modules may be implemented using hardware, one or more components or devices that execute software, or a combination thereof. For instance, one or more of the units or modules may be an integrated circuit, such as a programmed FPGA, a GPU, or an ASIC. It will be appreciated that where the modules are implemented using software for execution by a processor for example, they may be retrieved by a processor, in whole or part as needed, individually or together for processing, in single or multiple instances, and that the modules themselves may include instructions for further deployment and instantiation.

[00172] Additional details regarding the EDs 110, T-TRP 170, and NT-TRP 172 are known to those of skill in the art. As such, these details are omitted here.

[00173] FIG. 5 is a schematic diagram of an AI-based communication device.

[00174] A wireless system includes a plurality of connected devices. A device 500 is either base station (BS) or user equipment (UE). The device 500 may have three systems: sensing system 510, communication system 520, and/or AI system 530. The sensing system 510 senses and collects signals and data, the communication system 520 transmits and receives signals and data, and the AI system 530 trains and infers the AI implementations. An exemplary AI implementation is based on two cycles of deep learning, a training cycle and an inference cycle. In some possible application scenarios, the training cycle can also be referred to as the learning cycle and the inference cycle can also be referred to as the reasoning cycle.

[00175] Deep learning consists of two cycles: training (or learning) and inference (or reasoning). In a training cycle, the coefficients of neurons are learned from training data to fulfill a specific training goal or target. In the inference or reasoning cycle, an input data sample is fed into a trained neural network that would output a prediction.

[00176] During a training cycle, the AI system 530 of the device 500 may train the DNN or DNNs where the sensing system 510 of the device 500 may generate signals and/or data. The communication system 520 of the device 500 may receive the signals or data from another device or other devices. During and/or after the AI system 530 finishes training, the communication of the device may transmit the training results to another device or other devices.

[00177] During an inference cycle, the AI system 530 of a device 500 may perform one inference or a series of inferences with one DNN or DNNs to fulfill one task or tasks, where the sensing system 510 of the device 500 may generate signals and/or data, the communication system 520 of the device 500 may receive signals or data from another device or other devices. After the AI system 530 of the device 500 finishes inferencing, the communication system 520 of the device 500 may transmit the inferencing results to another device or other devices.

[00178] The AI implementations may either switch between the two cycles or stay in the two cycles simultaneously. For example, the AI system 530 of the device 500 may train the second DNN but still performs inference on the first DNN.

[00179] During the training cycle, the AI system 530 of the device 500 can work in single-user mode. In this mode, the

AI system 530 trains the DNN or DNN(s) with the data provided by the sensing system 510 of the device 500. Examples of the data include local sensing data and local channel data. Local sensing data includes RGB data, light detection and ranging (LiDAR) data, temperature data, air pressure data, electric outrage data, etc. Local channel data includes channel state information (CSI), received signal strength indicator (RSSI), latency data, etc.

[00180] Alternatively, the AI system 530 of the device 500 may work in a cooperative mode. In this mode, the AI system 530 trains the DNN or DNN(s) with the data that the communication system 520 of the device 500 receives. Example data includes sensing data, channel data, neuron data and latent output data. Sensing data includes RGB data, LiDAR data, temperature data, air pressure data, electric outrage data, etc. Channel data includes CSI, RSSI, delay data, etc. Neuron data includes a number of neurons or a number of gradients. Latent output data includes several latent outputs.

[00181] FIG. 6 is a schematic diagram of a device 500 receiving reference data samples from a device 600 according to an embodiment of the present application. The AI system 530 of the device 500 in cooperative mode may use data such as: accumulating the sensing data that the communication system 520 of the device 500 received into one training data set; accumulating the channel data that the communication system 520 of the device 500 received into one training data set; setting local neurons by the neurons that the communication system 520 of the device 500 received, which is a typical federated learning scheme; inputting the latent outputs that the communication system 520 of the device 500 received to its DNN(s).

[00182] Alternatively, the AI system 530 of the device 500 in a cooperative mode may use the data that the communication system 520 of the device 500 received together with its local ones, such as: mixing the local sensing data that the sensing system 510 of the device 500 provided with the sensing data that the communication system 520 of the device 500 received into one training data set; mixing the local channel data that the sensing system 510 of the device 500 provided with the channel data that the communication system 520 of the device 500 received into one training data set; averaging the local neurons that the AI system 530 of the device 500 possessed with the neurons that the communication system 520 of the device 500 received, which is a typical federated learning scheme; averaging the local latent outputs that the AI system 530 of the device 500 possessed and inputting them to its DNN(s).

[00183] FIG. 7 is a schematic diagram of reference data samples consisting of a plurality of groups according to an embodiment of the present application. During the training cycle, the communication system 520 of the device 500 may receive some reference data samples in both single-user or cooperative mode. Some devices transmit the reference data samples in broadcast, multicast, or unicast channels. The other devices transmits an indicator or indicators about which layer or layers to which the reference data samples are related, where, for example, there are three groups of the reference data samples: the first group of the reference data samples is indicated to be related to the input layer to the DNN, the second group of the reference data samples is indicated to be related to one latent layer output of the DNN, and the third group of the reference data samples

is indicated to be related to the layer output from the DNN.

[00184] The AI system 530 of the device 500 may measure the distances between its local data samples and reference data samples group by group. The AI system 530 of the device 500 may randomly, non-randomly, uniformly, or non-uniformly sample its local layer inputs, local latent layer outputs, and/or layer outputs. Then the AI system 530 of the device 500 measures the distance between the local samples and the reference samples that the communication system 520 of the device 500 received. If the average distances of all the groups are consistently below a predefined threshold or thresholds, the AI system 530 of the device 500 may tell that the current training procedure works as expected, otherwise the AI system 530 may tell it is abnormal.

[00185] In a case where a device has no AI system but has sensing and communication systems, the sensing system of the device may be still able to measure the distances between its local data sample(s) and the reference data sample(s) related to the layer input to the DNN. If the average distance on the layer input is below a predefined threshold, the sensing system of the device may consider that the sensing device is catching "good" data, otherwise bad data. The communication system of the device may transmit only good data to other devices and may not transmit bad data to other devices, or the communication system of the device may label the sensing data with the distance before transmitting them to other devices.

[00186] To protect raw data and save bandwidth, a group of the reference data samples are encoded or compressed to a lower dimensional space than their original space. The encoder or compressor can be linear or non-linear. A linear encoder can be realized with some standard basis such as Fourier Basis, discrete cosine transform (DCT), wavelets, or a linear encoder can be with some customized basis. These bases may consist of a unitary matrix (orthonormal). A non-linear encoder can be realized with some DNNs. FIG. 8 is a schematic representation of a DNN-based approximation according to an embodiment of the present application.

[00187] Unlike the traditional compression schemes built for reliable reconstruction, the encoder deliberately avoids a reliable reconstruction but preserves as much topological distances as possible, when the data is compressed into a lower dimensional space. That is, the relative distance between two data samples in their original signal space may be well preserved after being encoded into a low-dimensional space.

[00188] In wireless systems, an AI-based solution may be in a form of an auto-encoder (AE), whose encoding DNN is on the transmitter side and decoding DNN on the receiver side. The encoding DNN and decoding DNN are likely trained and provided by different providers. Moreover, as DNN is considered as highly intellectual property, it is hard for AI provider to open their DNN models. In this case, a wireless system can help interconnect the two.

[00189] FIG. 9 is a schematic diagram of an information bottleneck as a learning ratio according to an embodiment of the present application.

[00190] During the training cycle, the AI system 530 of a device 500 may work in a single user mode or cooperative

mode. In both modes the AI system 530 of the device 500 may calculate the learning metric or metrics over the time periods, epoch by epoch, or batch by batch. The learning metric or metrics may be as a function of the timing periods (epoch or batch), t, and/or one of the M latent layers $\mathbb{T}(t) = [T_1(t), T_2(t), \ldots, T_M(t)]$. The learning metric or metrics on a latent layer may include: $\delta_1(\mathbf{T}_m(t), \mathbf{X}(t))$, $\delta_2(\mathbf{T}_m(t), \mathbf{Y}(t))$, and/or $\rho_m(t) = \frac{\delta_1(\mathbf{T}_m(t), \mathbf{X}(t))}{\delta_2(\mathbf{T}_m(t), \mathbf{Y}(t))}$. $\delta_1(\mathbf{T}_m(t), \mathbf{X}(t))$ is a distance between the distribution of the input, $\mathbf{X}(t)$, to the DNN and the distribution of the m-th latent layer's output $\mathbf{T}_m(t)$ at the t-th epoch. $\delta_2(\mathbf{T}_m(t), \mathbf{Y}(t))$ is a distance between the distribution of the output, $\mathbf{Y}(t)$, from the DNN and the distribution of the m-th latent layer's output $\mathbf{T}_m(t)$ at the t-th epoch.

[00191] According to information bottleneck theory, if $\delta_1(\mathbf{T}_m(t), \mathbf{X}(t))$ and $\delta_2(\mathbf{T}_m(t), \mathbf{Y}(t))$ are the mutual information between $\mathbf{T}_m(t)$ and $\mathbf{X}(t)$ and the mutual information between $\mathbf{T}_m(t)$ and $\mathbf{Y}(t)$, $\delta_1(\mathbf{T}_m(t), \mathbf{X}(t))$ decreases over the layers: $\delta_1(\mathbf{T}_{m+1}(t), \mathbf{X}(t)) \leq \delta_1(\mathbf{T}_m(t), \mathbf{X}(t))$ and $\delta_2(\mathbf{T}_m(t), \mathbf{Y}(t))$ increase over the layers: $\delta_2(\mathbf{T}_{m+1}(t), \mathbf{Y}(t)) \geq \delta_2(\mathbf{T}_m(t), \mathbf{Y}(t))$. $\delta_1(\mathbf{T}_m(t), \mathbf{X}(t))$ decreases over the timing periods: $\delta_1(\mathbf{T}_m(t+1), \mathbf{X}(t+1)) \leq \delta_1(\mathbf{T}_m(t), \mathbf{X}(t))$ and $\delta_2(\mathbf{T}_m(t), \mathbf{Y}(t))$ increase over the timing periods: $\delta_2(\mathbf{T}_m(t+1), \mathbf{Y}(t+1)) \geq \delta_2(\mathbf{T}_m(t), \mathbf{Y}(t))$.

[00192] Therefore, if the learning cycle is normal, during a specific timing period t, $\rho_m(t)$ is decreasing: $\rho_1(t) > \rho_2(t) > \cdots > \rho_M(t)$. If the learning cycle is normal, on the m-th layer, $\rho_m(t)$ is decreasing: $\rho_m(t) > \rho_m(t+1) > \cdots > \rho_m(t+\Delta t)$. In practice, mutual information can be approximated by Hilbert-Schmidt independence criterion (HSIC), Jensen-Shannon divergence (JSD), Kullback-Leibler (KL), and so on. Basically, if a method to approximate mutual information doesn't change the tendencies above, it can be used as the method to compute the distance.

[00193] The communication system 520 of a device 500 receives a message that asks for measuring the learning metric ratio(s), which specifies on which layers in which period to measure which learning metric ratios in which method.

[00194] The AI system 530 of the device 500 may perform the measurements and computations on its DNN undertrained according to the message. The AI system 530 may store the learning metric ratios as a function of the layers and timing periods. The AI system 530 may do the statistics on the accumulated learning metric ratios to check if the learning metric ratios satisfy the decreasing or increasing properties above. If the AI system 530 of the device 500 suspects an abnormal decrease or increase of the learning metric ratios, it may choose to send an alarming message.

[00195] The communication system 520 of the device 500 may report the learning metric ratios in the requested periods according to the message, or the communication system 520 of the device 500 may report the learning metric ratios that the AI system 530 of the device 500 judges as abnormal.

[00196] FIG. 10 is a schematic diagram of autoencoders including the cross check on the learning metric ratios according to an embodiment of the present application.

[00197] A system consists of one device as the first device and another device as the second device. The AI system of the first device trains the first DNN-based autoencoder by its local data, and the AI system of the second device trains the second DNN-based autoencoder by its local data.

[00198] The communication system of the first device may send a message to the second device to ask the second device to measure and feedback the learning metric ratios.

[00199] The communication system of the second device may receive the message so that the AI system of the second device may perform the measurement and computations on its DNN undertrained according to the message. The AI system of the second device may store the learning metric ratios as a function of the layers and timing periods. The AI system of the second device may do the statistics on the accumulated learning metric ratios to check if the learning metric ratios satisfy the decreasing or increasing properties above. If the AI system of the second device suspects an abnormal decrease or increase of the learning metric ratios, it may choose to send an alarming message. The communication system of the second device may report the learning metric ratios in the requested periods to the first device according to the message, or the communication system of the second device may report the learning metric ratios that the AI system of the device judges as abnormal to the first device.

[00200] The first device may be an encoding device and the second device may be a decoding device, and the encoding DNN of the encoding device may be output to the decoding DNN of the decoding device. Alternatively, the first device may be a decoding device and the second device may be an encoding device, and the encoding DNN of the encoding device may be output to the decoding DNN of the decoding device.

[00201] In scenarios with a two-sided model (e.g., encoder and decoder), the BS uses an encoder model and the UE uses a decoder model (a similar solution for the case of a decoder model at the BS and an encoder model at the UE). The encoder model and decoder model should be matched, so they can be interoperable. The joint inference of the two-sided (AI/ML) model consists of AI/ML inference that is jointly performed by a UE and a network device, i.e., the first part of the inference is first performed by the UE, and then the remaining part is performed by the network device, and vice versa.

[00202] FIG. 11 is a schematic diagram of a split encoder and decoder according to an embodiment of the present application.

[00203] The devices jointly train the encoder and decoder, the encoder and decoder structure depends on the split point in the joint model of encoder and decoder. For example, the joint model is split into two parts, where one part is an encoder and the other part is a decoder. There are one or multiple candidate split points, where the split point is a latent layer in the joint model.

[00204] Optionally, the encoder and decoder can use the same neural network model or neural network structure, e.g.

the encoder uses 60 layers of it and the decoder uses the remaining 40 layers. Optionally, the encoder and decoder may also use separate neural network models.

[00205] A first AI model is an encoder and a second AI model is a decoder. Alternatively, the first AI model is a decoder and the second AI model is an encoder. The first AI model and the second AI model constitute a two-sided model.

[00206] In one possible scenario, the first AI model is at the UE side and the second AI model is at the BS side.

[00207] In another possible implementation scenario, the first AI model 1 and the second AI model 1 are at the UE side and the first AI model 2 and the second AI model 2 are at the BS side. For example, UE trains its own encoder1 and decoder1, BS trains its own encoder2 and decoder2, and finally aligns the UE's encoder 1 with the BS's decoder 2.

[00208] As shown in FIG. 11, the BS and UE train their encoder and decoder models individually, i.e., there are multiple sets of {encoders, decoders} at the BS side and multiple sets of {encoders, decoders} at the UE side. Depending on the split point 1 of the BS side model, the encoder model of the BS can be different. Similarly depending on the split point 2 of the UE side model, the decoder model of the UE can be different.

[00209] FIG. 12 is a schematic diagram of an auto-encoder model according to an embodiment of the present application.

[00210] In a two-sided model, if the encoder is deployed at a user equipment (UE) side and the decoder is deployed at a base station (BS) side, X is the input of the encoder, Y is the output of the decoder, and T is the output of the encoder as well as the input of the decoder. T is exchanged through the air interface between the BS and the UE.

[00211] In a two-sided model, if encoder 1 and decoder 1 are deployed at the UE side, X is the input of the encoder 1, Y is the output of the decoder 1, and T is the output of the encoder 1 and also the input of the decoder 1. T can also be regarded as a latent layer or split point of the joint model of encoder 1 and decoder 1.

[00212] In a two-sided model, if encoder 2 and decoder 2 are deployed at the BS side, X is the input of the encoder 2, Y is the output of the decoder 2, and T is the output of the encoder 2 and also the input of the decoder 2. T can also be regarded as a latent layer or split point of the joint model of encoder 2 and decoder 2.

[00213] The BS and UE train their models independently and need to align the encoder and decoder. In addition, during inference, the generalization problem at UE or BS needs to be considered. For example, the generalization performance of the paired encoder from UE and decoder from BS becomes worse, but it is hard to know whether the current encoder model is outdated or not during the inference process.

[00214] In the following embodiments, the method provided by the present application is described in terms of the BS using an encoder model and the UE using a decoder. For the case where the BS uses the decoder model and the UE uses the encoder model, it is a similar solution, which will not be repeated in this application.

[00215] In embodiments of the present application, the BS and the UE can interoperate with each other by some

configured or predefined rules, e.g., ratios between layer outputs and model inputs, ratios between layer outputs and model outputs.

[00216] The BS can send a mutual information approximation method to the UE, such as Hilbert-Schmidt Independence Criterion (HSIC), or a predefined mutual information approximation method.

[00217] A theory of information bottleneck can be used to study AI/ML model. Consider X and Y respectively as input and output layers of a DNN, and let T be any hidden layer of the network. Let I(X,T) denote the amount of information that the hidden layer contains about the input and let I(T,Y) denote the amount of information that the hidden layer contains about the output. During the training process, the I(X,T) should be decreased and I(T,Y) should be increased. For the latter layer which is near the output layer, the finalized I(T,Y) should be larger, e.g. for the last layer, I(T,Y) should be equal to Y ideally.

[00218] FIG. 13 is a flowchart of a communication method according to an embodiment of the present application.

[00219] 610, sending first information.

[00220] The first information is the information related to the mutual information of a first AI model and a second AI model. The first AI model and the second AI model constitute a two-sided model. A first AI model is an encoder and a second AI model is a decoder. Alternatively, a first AI model is a decoder and a second AI model is an encoder.

[00221] In one possible scenario, the first AI model is at the UE side and the second AI model is at the BS side.

[00222] In another possible implementation scenario, the first AI model 1 and the second AI model 1 are at the UE side and the first AI model 2 and the second AI model 2 are at the BS side. For example, UE trains its own encoder1 and decoder1, BS trains its own encoder2 and decoder2, and finally aligns the UE's encoder 1 with the BS's decoder 2. In this case, first information is related to the mutual information of encoder1 and decoder1.

[00223] 620, receiving a first message.

[00224] The first message indicates the AI model associated with the first AI model. The BS can determine whether the first message is within the predetermined range, and if it is outside the predetermined range, the BS can send a first message indicating the UE to switch the AI model, or indicating the UE to use the specified AI model.

[00225] FIG. 14 is a flowchart of a communication method according to an embodiment of the present application.

[00226] 710, BS configures a ratio or ratio range or mutual information or neuron node size of a layer.

[00227] The layer that the BS configures can be a latent layer of a model, or an input layer of a model, or an output layer of a model.

[00228] The BS can configure the ratio or ratio range or mutual information (function-1(I(X, T)) and/or function-2(I(T, Y))) or neuron node size of the layer, where the layer or the output of the encoder or the input of the decoder can be indicated.

[00229] 720, UE determines the mutual information related information.

[00230] The UE can calculate a ratio or a ratio range of mutual information A and mutual information B, where the mutual information A is I(X,T) and the mutual information B is I(T,Y). The UE can calculate the ratio of the mutual information A and the mutual information B. Alternatively, the UE can calculate a ratio range of the mutual information A and the mutual information B. Alternatively, the UE can calculate the value(s) of the mutual information A and/or the mutual information B. Alternatively, the UE can determine a neuron network node size, which is configured to indicate an output format of the first AI model or an input format of the second AI model.

[00231] FIG. 15 is a schematic diagram of the UE calculating mutual information or mutual information ratios according to an embodiment of the present application. $T = f_1(X,\theta)$ denotes the relationship between T and X, with θ as a parameter. $Y = g_1(T,\varphi)$ denotes the relationship between T and Y, with φ as a parameter. In one possible implementation, the UE calculates the ratio of the mutual information A and the mutual information B. The ratio can be defined as equation (1), where function-1 and function-2 can be one of max(), min(), average() and so on. $\alpha(T)$ denotes the ratio of mutual information for the specified T layer. For example, $\alpha(T) = \frac{\max(I(T,Y))}{\min(I(X,T))}$. The function max() is configured to obtain the maximum value, the function min() is configured to obtain the minimum value, and the function average() is configured to obtain the average value.

[00232] $$\alpha(T) = \frac{\text{function}-2(I(T,Y))}{\text{function}-1(I(X,T))} \quad (1)$$

[00233] In one possible implementation, the UE calculates the value(s) of the mutual information A and/or the mutual information B. The value of the mutual information A is function-1(I(X, T)), and the value of the mutual information B is function-2(I(T, Y)). The function-1 and function-2 can be one of max(), min(), average() and so on.

[00234] The ratio range is a range of multiple ratios. The ratio range can also be some discrete value, such as a collection of values.

[00235] 730, UE determines which decoder shall be used.

[00236] According to the configured ratio or ratio range or mutual information (function-1(I(X, T)) and/or function-2(I(T, Y))) or neuron node size, UE can determine which decoder shall be used, e.g. determine the split point between its trained encoder and decoder. The UE can report the used decoder to the BS.

[00237] In a possible implementation, if the ratio or ratio range or mutual information or neuron node size of the UE is out of range, the UE can send a model switching request to the BS. The BS receives the model switching request and sends a message indicating the UE to perform the switching of the model.

[00238] By the method provided by the embodiments of the present application, the BS can operate the model of the UE. the method for the UE to operate the BS is the same, and will not be repeated in the present application. The methods provided by the embodiments of the present application can realize interoperability between the AI model of the BS and the AI model of the UE.

[00239] FIG. 16 is a flowchart of a communication method according to an embodiment of the present application. In one possible embodiment, during the inference process, the UE would keep reporting its ratio or ratio range or mutual information or neuron node size on selected layer(s). If the ratio stays within some pre-defined range, the AI model at the UE side is suitable. Otherwise, the BS asks the UE to switch the AI model.

[00240] 810, BS configures the ratio or ratio range or mutual information or neuron node size of a layer.

[00241] The layer that the BS configures can be a latent layer of a model, or an input layer of a model, or an output layer of a model.

[00242] The BS can configure the ratio or ratio range or mutual information (function-1(I(X, T)) and/or function-2(I(T, Y))) or neuron node size of the layer, where the layer or the output of the encoder or the input of the decoder can be indicated.

[00243] 820, UE reports its ratio or ratio range or mutual information or neuron node size of the layer.

[00244] The UE can calculate a ratio or a ratio range of mutual information A and mutual information B, where the mutual information A is I(X,T) and the mutual information B is I(T,Y). The UE can calculate the ratio of the mutual information A and the mutual information B. Alternatively, the UE can calculate a ratio range of the mutual information A and the mutual information B. Alternatively, the UE can calculate the value(s) of the mutual information A and/or the mutual information B. Alternatively, the UE can determine a neuron network node size, which is configured to indicate an output format of the first AI model or an input format of the second AI model.

[00245] FIG. 15 is a schematic diagram of the UE calculating mutual information or mutual information ratios according to an embodiment of the present application. $\mathrm{T} = f_1(\mathrm{X},\theta)$ denotes the relationship between T and X, with θ as a parameter. $\mathrm{Y} = g_1(\mathrm{T},\varphi)$ denotes the relationship between T and Y, with φ as a parameter. In one possible implementation, the UE calculates the ratio of the mutual information A and the mutual information B. The ratio can be defined as equation (1), where function-1 and function-2 can be one of max(), min(), average(), and so on. For example, $\alpha(\mathrm{T}) = \frac{\max\left(\mathrm{I}(\mathrm{T},\mathrm{Y})\right)}{\min\left(\mathrm{I}(\mathrm{X},\mathrm{T})\right)}$. $\alpha(\mathrm{T})$ denotes the ratio of mutual information for the specified T layer.

[00246] In one possible implementation, the UE calculates the value(s) of the mutual information A and/or the mutual information B. The value of the mutual information A is function-1(I(X, T)), and the value of the mutual information B is

function-2(I(T, Y)). The function-1 and function-2 can be one of max(), min(), average(), and so on.

[00247] The ratio range is a range of multiple ratios. The ratio range can also be some discrete value, such as a collection of values.

[00248] The UE can report to the BS its ratio or ratio range or mutual information of the layer. The mutual information of the layer includes (function-1(I(X, T)) and/or function-2(I(T, Y))).

[00249] The UE reports its ratio or ratio range or mutual information or neuron node size that can be periodical, semi-persistent, or aperiodic.

[00250] The reporting of the UE can be event-triggered reporting, e.g., the UE reports when its ratio or ratio range or mutual information or neuron node size changes.

[00251] 830, BS indicates the AI model to the UE.

[00252] When the ratio or ratio range or mutual information or neuron node size sent by the UE is out of range, the BS can send a message to the UE indicating the UE to switch an DNN model.

[00253] In a possible implementation, the UE can also send more than one of the ratio, the ratio range, the mutual information, and the neuron node size. The BS indicates the UE to switch a model when any of them is out of the range configured by the BS. Alternatively, the BS can indicate the UE to use a specific AI model.

[00254] By the method provided by the embodiments of the present application, the BS can operate the model of the UE. the method for the UE to operate the BS is the same, and will not be repeated in the present application. The methods provided by the embodiments of the present application can realize interoperability between the AI model of the BS and the AI model of the UE.

[00255] FIG. 17 is a flowchart of a communication method according to an embodiment of the present application.

[00256] 910, BS configures multiple mutual information ratio ranges.

[00257] The BS configures multiple mutual information ratio ranges (a set of $\alpha(T)$). Each ratio range is configured with a corresponding ratio range index.

[00258] In one possible implementation, a ratio range index (1~N) indicates a ratio range, e.g. x%~y%. For example, index 1 corresponds to x1%~y1% and index 2 corresponds to x2%~y2%. The ratio range can also be some discrete value, such as a collection of values.

[00259] 920, UE reports which ratio range index is aligned and chosen as its model reference range.

[00260] In embodiments of the present application, the UE can calculate a ratio or a ratio range of mutual information A and mutual information B, where the mutual information A is I(X,T) and the mutual information B is I(T,Y). The UE can calculate the ratio of the mutual information A and the mutual information B. Alternatively, the UE can calculate a ratio range

of the mutual information A and the mutual information B.

[00261] In one possible implementation, the UE calculates the ratio of the mutual information A and the mutual information B. The ratio can be defined as equation (1), where function-1 and function-2 can be one of max(), min(), average() and so on. $\alpha(\mathrm{T})$ denotes the ratio of mutual information for the specified T layer. For example, $\alpha(\mathrm{T})=\frac{\max\left(\mathrm{I}(\mathrm{T},\mathrm{Y})\right)}{\min\left(\mathrm{I}(\mathrm{X},\mathrm{T})\right)}$.

[00262] The UE determines which configured ratio range its ratio is aligned to. The UE reports the index of the aligned rate range and selects that rate range as the rate reference range for its own model.

[00263] 930, BS uses appropriate model to interoperate with the UE's model.

[00264] On receiving the index of the report at the BS, the BS knows the features of the model selected by the UE. The features include a ratio range of the mutual information I(X, T) and/or I(T, Y), so that the BS can use appropriate model to interoperate with the UE's model. Alternatively, the BS can also indicate the UE to use the specific AI model to interoperate with the AI model at the BS side.

[00265] Similarly, the BS can configure multiple ratios, pieces of mutual information I(X, T), pieces of mutual information I(T, Y), or neuron network node sizes. Each ratio, mutual information I(X, T), mutual information I(T, Y) or neural network node size corresponds to an index. The UE aligns its AI model with the AI model at the BS side by reporting an index corresponding to the ratio, an index corresponding to the mutual information I(X, T), an index corresponding to the mutual information I(T, Y), or an index corresponding to the neuron network node size. Optionally, one index can also correspond to a combination of any of the following: ratios, ratio ranges, mutual information, and neuron node sizes. For example, index 1 corresponds to a particular ratio range and a neuron node size. The above embodiments should not be construed as a limitation of the present application.

[00266] By the method provided by the embodiments of the present application, the BS can operate the model of the UE. the method for the UE to operate the BS is the same, and will not be repeated in the present application. The methods provided by the embodiments of the present application can realize interoperability between the AI model of the BS and the AI model of the UE.

[00267] FIG. 18 is a schematic block diagram of a communication apparatus 1800 according to an embodiment of this application. The communication apparatus 1800 includes: a sending module 1810 configured to send first information related to mutual information of a first artificial intelligence (AI) model and a second AI model, the first AI model and the second AI model constituting a two-sided model; and a receiving module 1820 configured to receive a first message indicating an AI model related to the first AI model.

[00268] In a possible implementation, the first information includes at least one of first mutual information, second

mutual information, a first ratio, a first neuron node size, and a first ratio range.

[00269] In a possible implementation, the first information includes an index corresponding to first mutual information, an index corresponding to second mutual information, an index corresponding to a first ratio, an index corresponding to a first neuron node size, or an index corresponding to a first ratio range.

[00270] In a possible implementation, at least one piece of third mutual information, at least one piece of fourth mutual information, at least one second ratio, at least one second neuron node size or at least one second ratio range is predetermined or configured by a network device, where each of third mutual information, fourth mutual information, second ratios, second neuron node sizes or second ratio ranges corresponds to an index, where the first mutual information is one of the at least one piece of third mutual information, the second mutual information is one of the at least one piece of fourth mutual information, the first ratio is one of the at least one second ratio, the first neuron node size is one of the at least one second neuron node size, and the first ratio range is one of the at least one second ratio range.

[00271] In a possible implementation, the first mutual information is an amount of information about an input included in an output of the first AI model, the second mutual information is an amount of information about an output included in an input of the second AI model, the first ratio is a ratio of the second mutual information to the first mutual information, the first neuron node size is configured to indicate an output format of the first AI model or an input format of the second AI model, and the first ratio range is a range of multiple first ratios.

[00272] In a possible implementation, the sending module 1810 is further configured to send the first information when the first information has changed during a time period.

[00273] In a possible implementation, the sending module 1810 is further configured to send a model switch request.

[00274] In a possible implementation, the receiving module 1820 is further configured to receive a second message indicating a method for calculating the first information.

[00275] In a possible implementation, the method for calculating the first information is Hilbert-Schmidt independence criterion (HSIC), or a predefined mutual information approximation method.

[00276] In a possible implementation, the first AI model is a decoder and the second AI model is an encoder; or, the first AI model is an encoder and the second AI model is a decoder.

[00277] In a possible implementation, the first AI model is at a user equipment side and the second AI model is at a network device side; or the first AI model and the second AI model are at a user equipment side.

[00278] In a possible implementation, the apparatus is located on a user equipment.

[00279] FIG. 19 is a schematic block diagram of a communication apparatus 1900 according to an embodiment of this application. The communication apparatus 1900 includes: a receiving module 1910 configured to receive first information

related to mutual information of a first AI model and a second AI model, the first AI model and the second AI model constituting a two-sided model; and a sending module 1920 configured to send a first message indicating an AI model related to the first AI model.

[00280] In a possible implementation, the first information includes at least one of first mutual information, second mutual information, a first ratio, a first neuron node size, and a first ratio range.

[00281] In a possible implementation, the first information includes an index corresponding to first mutual information, an index corresponding to second mutual information, an index corresponding to a first ratio, an index corresponding to a first neuron node size, or an index corresponding to a first ratio range.

[00282] In a possible implementation, at least one piece of third mutual information, at least one piece of fourth mutual information, at least one second ratio, at least one second neuron node size or at least one second ratio range is predetermined or configured by a network device, where each of third mutual information, fourth mutual information, second ratios, second neuron node sizes or second ratio ranges corresponds to an index, where the first mutual information is one of the at least one piece of third mutual information, the second mutual information is one of the at least one piece of fourth mutual information, the first ratio is one of the at least one second ratio, the first neuron node size is one of the at least one second neuron node size, and the first ratio range is one of the at least one second ratio range.

[00283] In a possible implementation, the first mutual information is an amount of information about an input piece of in an output of the first AI model, the second mutual information is an amount of information about an output piece of in an input of the second AI model, the first ratio is a ratio of the second mutual information to the first mutual information, the first neuron node size is configured to indicate an output format of the first AI model or an input format of the second AI model, and the first ratio range is a range of multiple first ratios.

[00284] In a possible implementation, the sending module 1920 is further configured to send the first information when the first information has changed during a time period.

[00285] In a possible implementation, the sending module 1920 is further configured to send a model switch request.

[00286] In a possible implementation, the receiving module 1910 is further configured to receive a second message indicating a method for calculating the first information.

[00287] In a possible implementation, the method for calculating the first information is Hilbert-Schmidt independence criterion (HSIC), or a predefined mutual information approximation method.

[00288] In a possible implementation, the first AI model is a decoder and the second AI model is an encoder; or, the first AI model is an encoder and the second AI model is a decoder.

[00289] In a possible implementation, the first AI model is at a user equipment side and the second AI model is at a

network device side; or the first AI model and the second AI model are at a user equipment side.

[00290] In a possible implementation, the apparatus is located on a user equipment.

[00291] As shown in FIG. 20, a communication apparatus 2000 can include a processor 2010 and a transceiver 2020. Optionally, the communication apparatus 2000 can further include a memory 2030. The memory 2030 can be configured to store indication information, or can be configured to store code, instructions, and the like that is to be executed by the processor 2010.

[00292] The memory 2030 can include a random memory, a flash memory, a read-only memory, a programmable read-only memory, a non-volatile memory, a register, or the like. The processor 2010 can be a central processing unit (CPU).

[00293] For other functions and operations of the communication apparatus 2000, refer to processes of the method embodiments from FIG. 5 to FIG. 17, which are not described again herein to avoid repetition.

[00294] An embodiment of the present application further provides a computer storage medium, and the computer storage medium can store a program instruction for performing the steps in the foregoing methods.

[00295] Optionally, the storage medium can be specifically the memory 2030.

[00296] An embodiment of the present application further provides a computer program product. The computer program product includes computer program code. When the computer program code runs on a computer, the computer is enabled to perform the steps in the foregoing methods.

[00297] Optionally, all or a part of computer program code can be stored in on a first storage medium. The first storage medium can be packaged together with the processor or separately with the processor.

[00298] An embodiment of the present application further provides a chip system, where the chip system includes an input/output interface, at least one processor, at least one memory, and a bus. The at least one memory is configured to store instructions, and the at least one processor is configured to invoke the instructions of the at least one memory to perform operations in the methods in the foregoing embodiments.

[00299] A person of ordinary skill in the art may understand that all or some of the processes of the methods in the embodiments may be implemented by a computer program instructing related hardware. The program may be stored in a computer-readable storage medium. When the program runs, the processes of the methods in the embodiments are performed. The foregoing storage medium may include: a magnetic disk, an optical disc, a read-only memory (ROM), or a random-access memory (RAM).

[00300] In the several embodiments provided in the present application, it should be understood that the disclosed system, apparatus, and method may be implemented in other manners. For example, the described apparatus embodiment is merely exemplary. For example, the unit division is merely logical function division and may be other division in actual implementation.

For example, a plurality of units or components may be combined or integrated into another system, or some features may be ignored or not performed. In addition, the displayed or discussed mutual couplings or direct couplings or communication connections may be implemented by using some interfaces. The indirect couplings or communication connections between the apparatuses or units may be implemented in electronic, mechanical, or other forms.

[00301] 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 according to actual needs to achieve the objectives of the solutions of the embodiments.

[00302] In addition, functional units in the embodiments of the present invention may be integrated into one processing unit, or each of the units may exist alone physically, or two or more units are integrated into one unit.

[00303] The foregoing are merely exemplary embodiments of the present invention. A person skilled in the art may make various modifications and variations to the present invention without departing from and scope of the present invention.

CLAIMS

What is claimed is:

1. A communication method, comprising:

sending first information related to mutual information of a first artificial intelligence (AI) model and a second AI model, the first AI model and the second AI model constituting a two-sided model; and

receiving a first message indicating an AI model related to the first AI model.

2. The method according to claim 1, wherein the first information comprises at least one of first mutual information, second mutual information, a first ratio, a first neuron node size, and a first ratio range.

3. The method according to claim 1, wherein the first information comprises an index corresponding to first mutual information, an index corresponding to second mutual information, an index corresponding to a first ratio, an index corresponding to a first neuron node size, or an index corresponding to a first ratio range.

4. The method according to claim 3, wherein at least one piece of third mutual information, at least one piece of fourth mutual information, at least one second ratio, at least one second neuron node size or at least one second ratio range is predetermined or configured by a network device, wherein each of third mutual information, fourth mutual information, second ratios, second neuron node sizes or second ratio ranges corresponds to an index, wherein the first mutual information is one of the at least one piece of third mutual information, the second mutual information is one of the at least one piece of fourth mutual information, the first ratio is one of the at least one second ratio, the first neuron node size is one of the at least one second neuron node size, and the first ratio range is one of the at least one second ratio range.

5. The method according to any one of claims 2 to 4, wherein the first mutual information is an amount of information about an input included in an output of the first AI model, the second mutual information is an amount of information about an output included in an input of the second AI model, the first ratio is a ratio of the second mutual information to the first mutual information, the first neuron node size is configured to indicate an output format of the first AI model or an input format of the second AI model, and the first ratio range is a range of multiple first ratios.

6. The method according to any one of claims 1 to 5, wherein the sending first information, comprises:

sending the first information when the first information has changed during a time period.

7. The method according to any one of claims 1 to 5, wherein the sending first information, comprises:

sending a model switch request.

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

receiving a second message indicating a method for calculating the first information.

9. The method according to claim 8, wherein the method for calculating the first information is Hilbert-Schmidt independence criterion (HSIC), or a predefined mutual information approximation method.

10. The method according to any one of claims 1 to 9, wherein the first AI model is a decoder and the second AI model is an encoder; or, the first AI model is an encoder and the second AI model is a decoder.

11. The method according to any one of claims 1 to 10, wherein the first AI model is at a user equipment side and the second AI model is at a network device side; or

the first AI model and the second AI model are at a user equipment side.

12. The method according to any one of claims 1 to 11, wherein the method is executed by a user equipment.

13. A communication method, comprising:

receiving first information related to mutual information of a first artificial intelligence (AI) model and a second AI model, the first AI model and the second AI model constituting a two-sided model; and

sending a first message indicating an AI model related to the first AI model.

14. The method according to claim 13, wherein the first information comprises at least one of first mutual information, second mutual information, a first ratio, a first neuron node size, and a first ratio range.

15. The method according to claim 13, wherein the first information comprises an index corresponding to first mutual information, an index corresponding to second mutual information, an index corresponding to a first ratio, an index corresponding to a first neuron node size, or an index corresponding to a first ratio range.

16. The method according to claim 15, wherein at least one piece of third mutual information, at least one piece of fourth mutual information, at least one second ratio, at least one second neuron node size or at least one second ratio range is predetermined or configured by a network device, wherein each of third mutual information, fourth mutual information, second ratios, second neuron node sizes or second ratio ranges corresponds to an index, wherein the first mutual information is one of the at least one piece of third mutual information, the second mutual information is one of the at least one piece of fourth mutual information, the first ratio is one of the at least one second ratio, the first neuron node size is one of the at least one second neuron node size, and the first ratio range is one of the at least one second ratio range.

17. The method according to any one of claims 14 to 16, wherein the first mutual information is an amount of information about an input piece of in an output of the first AI model, the second mutual information is an amount of information about an output piece of in an input of the second AI model, the first ratio is a ratio of the second mutual information to the first mutual information, the first neuron node size is configured to indicate an output format of the first AI model or an input format of the second AI model, and the first ratio range is a range of multiple first ratios.

18. The method according to any one of claims 13 to 17, wherein the receiving first information, comprises:

receiving a model switch request.

19. The method according to any one of claims 13 to 18, further comprising:

sending a second message indicating a method for calculating the first information.

20. The method according to claim 19, wherein the method for calculating the first information is Hilbert-Schmidt independence criterion (HSIC), or a predefined mutual information approximation method.

21. The method according to any one of claims 13 to 20, wherein the sending a first message, comprises:

sending the first message when a value of the first information is not within a corresponding predetermined range.

22. The method according to any one of claims 13 to 21, further comprising:

adjusting the AI model at a network device side based on the first information.

23. The method according to any one of claims 13 to 22, wherein the first AI model is a decoder and the second AI model is an encoder; or, the first AI model is an encoder and the second AI model is a decoder.

24. The method according to any one of claims 13 to 23, wherein the first AI model is at a user equipment side and the second AI model is at a network device side; or

the first AI model and the second AI model are at a user equipment side.

25. The method according to any one of claims 13 to 24, wherein the method is executed by a network device.

26. A communication apparatus, comprising:

a sending module configured to send first information related to mutual information of a first artificial intelligence (AI) model and a second AI model, the first AI model and the second AI model constituting a two-sided model; and

a receiving module configured to receive a first message indicating an AI model related to the first AI model.

27. The communication apparatus according to claim 26, wherein the first information comprises at least one of first mutual information, second mutual information, a first ratio, a first neuron node size, and a first ratio range.

28. The communication apparatus according to claim 26, wherein the first information comprises an index corresponding to first mutual information, an index corresponding to second mutual information, an index corresponding to a first ratio, an index corresponding to a first neuron node size, or an index corresponding to a first ratio range.

29. The communication apparatus according to claim 28, wherein at least one piece of third mutual information, at least one piece of fourth mutual information, at least one second ratio, at least one second neuron node size or at least one second ratio range is predetermined or configured by a network device, wherein each of third mutual information, fourth mutual information, second ratios, second neuron node sizes or second ratio ranges corresponds to an index, wherein the first mutual information is one of the at least one piece of third mutual information, the second mutual information is one of the at least one

piece of fourth mutual information, the first ratio is one of the at least one second ratio, the first neuron node size is one of the at least one second neuron node size, and the first ratio range is one of the at least one second ratio range.

30. The communication apparatus according to any one of claims 27 to 29, wherein the first mutual information is an amount of information about an input included in an output of the first AI model, the second mutual information is an amount of information about an output included in an input of the second AI model, the first ratio is a ratio of the second mutual information to the first mutual information, the first neuron node size is configured to indicate an output format of the first AI model or an input format of the second AI model, and the first ratio range is a range of multiple first ratios.

31. The communication apparatus according to any one of claims 26 to 30, wherein the sending module is further configured to send the first information when the first information has changed during a time period.

32. The communication apparatus according to any one of claims 26 to 30, wherein the sending module is further configured to send a model switch request.

33. The communication apparatus according to any one of claims 26 to 32, wherein the receiving module is further configured to receive a second message indicating a method for calculating the first information.

34. The communication apparatus according to claim 33, wherein the method for calculating the first information is Hilbert-Schmidt independence criterion (HSIC), or a predefined mutual information approximation method.

35. The communication apparatus according to any one of claims 26 to 34, wherein the first AI model is a decoder and the second AI model is an encoder; or, the first AI model is an encoder and the second AI model is a decoder.

36. The communication apparatus according to any one of claims 26 to 35, wherein the first AI model is at a user equipment side and the second AI model is at a network device side; or

the first AI model and the second AI model are at a user equipment side.

37. The communication apparatus according to any one of claims 26 to 36, wherein the apparatus is located on a user equipment.

38. A communication apparatus, comprising:

a receiving module configured to receive first information related to mutual information of a first artificial intelligence (AI) model and a second AI model, the first AI model and the second AI model constituting a two-sided model; and

a sending module configured to send a first message indicating an AI model related to the first AI model.

39. The communication apparatus according to claim 38, wherein the first information comprises at least one of first mutual information, second mutual information, a first ratio, a first neuron node size, and a first ratio range.

40. The communication apparatus according to claim 38, wherein the first information comprises an index corresponding to first mutual information, an index corresponding to second mutual information, an index corresponding to a first ratio, an

index corresponding to a first neuron node size, or an index corresponding to a first ratio range.

41. The communication apparatus according to claim 40, wherein at least one piece of third mutual information, at least one piece of fourth mutual information, at least one second ratio, at least one second neuron node size or at least one second ratio range is predetermined or configured by a network device, wherein each of third mutual information, fourth mutual information, second ratios, second neuron node sizes or second ratio ranges corresponds to an index, wherein the first mutual information is one of the at least one piece of third mutual information, the second mutual information is one of the at least one piece of fourth mutual information, the first ratio is one of the at least one second ratio, the first neuron node size is one of the at least one second neuron node size, and the first ratio range is one of the at least one second ratio range.

42. The communication apparatus according to any one of claims 39 to 41, wherein the first mutual information is an amount of information about an input piece of in an output of the first AI model, the second mutual information is an amount of information about an output piece of in an input of the second AI model, the first ratio is a ratio of the second mutual information to the first mutual information, the first neuron node size is configured to indicate an output format of the first AI model or an input format of the second AI model, and the first ratio range is a range of multiple first ratios.

43. The communication apparatus according to any one of claims 38 to 42, wherein the receiving module is further configured to receive a model switch request.

44. The communication apparatus according to any one of claims 38 to 43, wherein the sending module is further configured to send a second message indicating a method for calculating the first information.

45. The communication apparatus according to claim 44, wherein the method for calculating the first information is Hilbert-Schmidt independence criterion (HSIC), or a predefined mutual information approximation method.

46. The communication apparatus according to any one of claims 38 to 45, wherein the sending module is further configured to send the first message when a value of the first information is not within a corresponding predetermined range.

47. The communication apparatus according to any one of claims 38 to 46, further comprising a processing module configured to adjust the AI model at a network device side based on the first information.

48. The communication apparatus according to any one of claims 38 to 47, wherein the first AI model is a decoder and the second AI model is an encoder; or, the first AI model is an encoder and the second AI model is a decoder.

49. The communication apparatus according to any one of claims 38 to 48, wherein the first AI model is at a user equipment side and the second AI model is at a network device side; or

the first AI model and the second AI model are at a user equipment side.

50. The communication apparatus according to any one of claims 38 to 49, wherein the apparatus is located on a network device.

51. A communication apparatus, comprising a processor and a memory, and the processor is connected to the memory; wherein the memory is configured to store instructions, and the processor is configured to execute the instructions; and when the processor executes the instructions stored in the memory, the processor is enabled to perform the method according to any one of claims 1 to 12, or the method according to any one of claims 13 to 25.

52. A communication system, comprising the communication apparatus according to any one of claims 26 to 37, and the communication apparatus according to any one of claims 38 to 50.

53. A computer-readable storage medium, wherein the computer-readable storage medium stores instructions, and when the instructions run on a processor, the processor is enabled to perform the method according to any one of claims 1 to 12, or the method according to any one of claims 13 to 25.

54. A computer program product, comprising computer program code, and when the computer program code runs on a computer, the computer is enabled to perform the method according to any one of claims 1 to 12, or the method according to any one of claims 13 to 25.

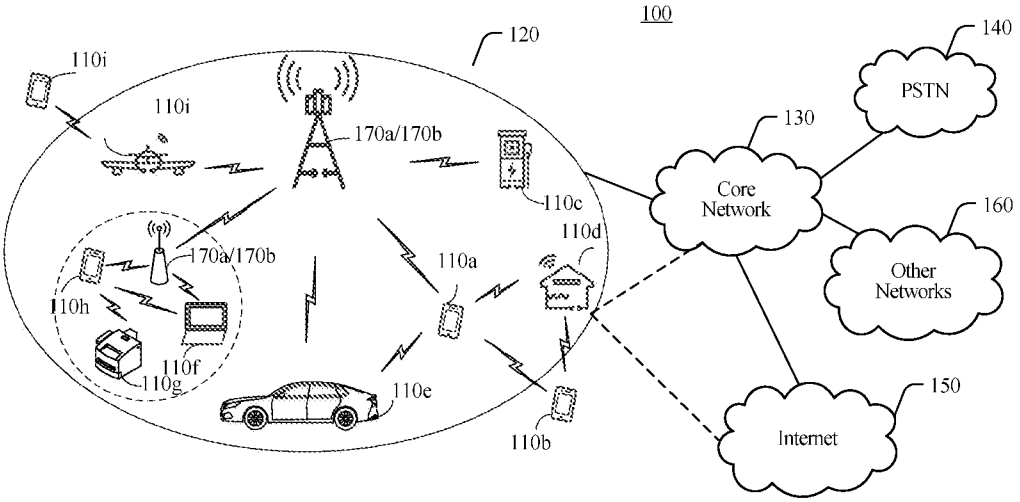

100
120
140
PSTN
110i
110i
170a/170b
130
Core Network
110c
110d
160
Other Networks
170a/170b
110a
110h
110f
110g
110e
150
Internet
110b

FIG. 1

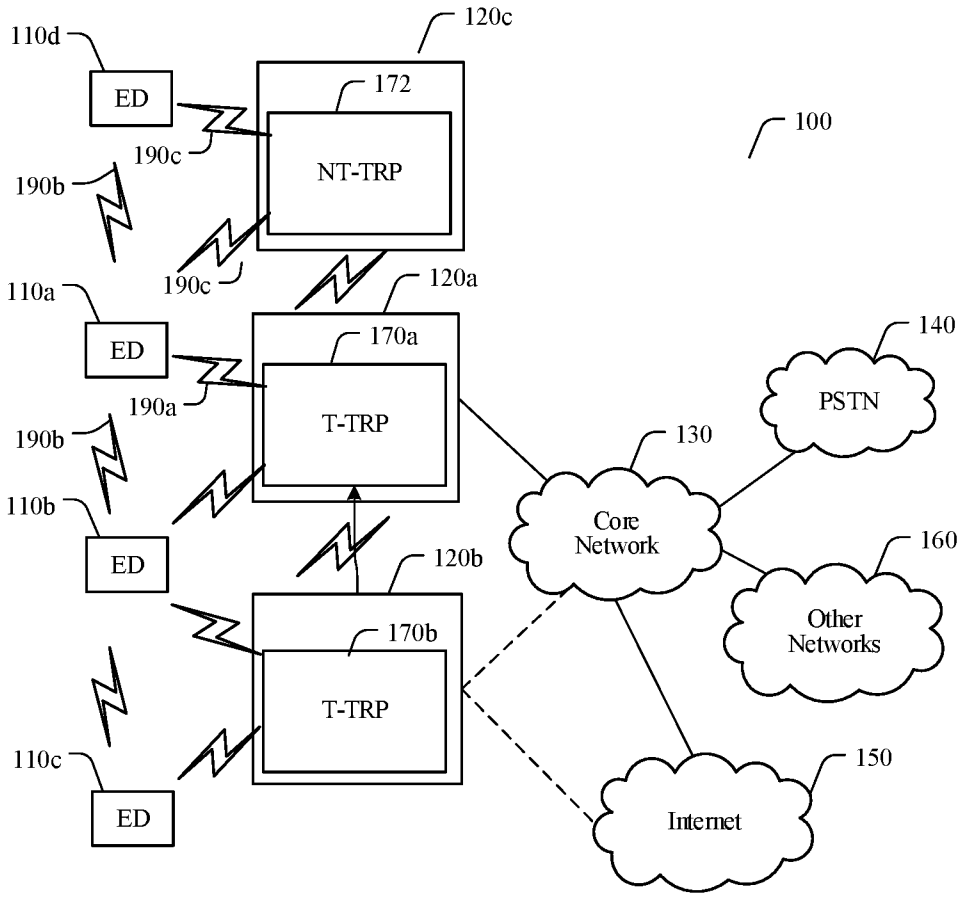

120c
110d
ED
172
100
190c
NT-TRP
190b
190c
120a
110a
ED
170a
140
PSTN
190a
T-TRP
130
190b
110b
Core Network
160
ED
120b
Other Networks
170b
T-TRP
110c
150
ED
Internet

FIG. 2

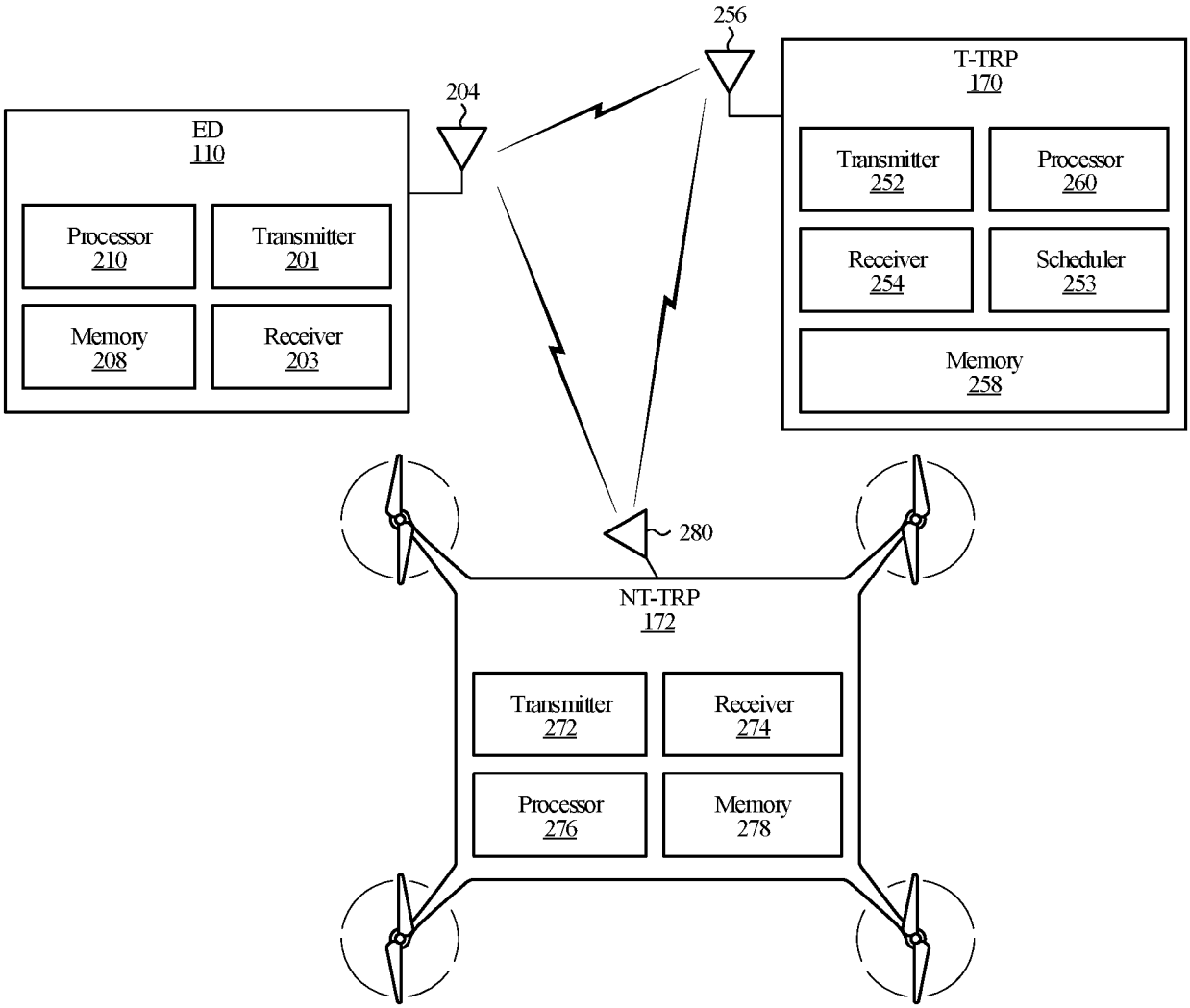

256
T-TRP
170
Transmitter
252
Processor
260
Receiver
254
Scheduler
253
Memory
258
204
ED
110
Processor
210
Transmitter
201
Memory
208
Receiver
203
280
NT-TRP
172
Transmitter
272
Receiver
274
Processor
276
Memory
278

FIG.3

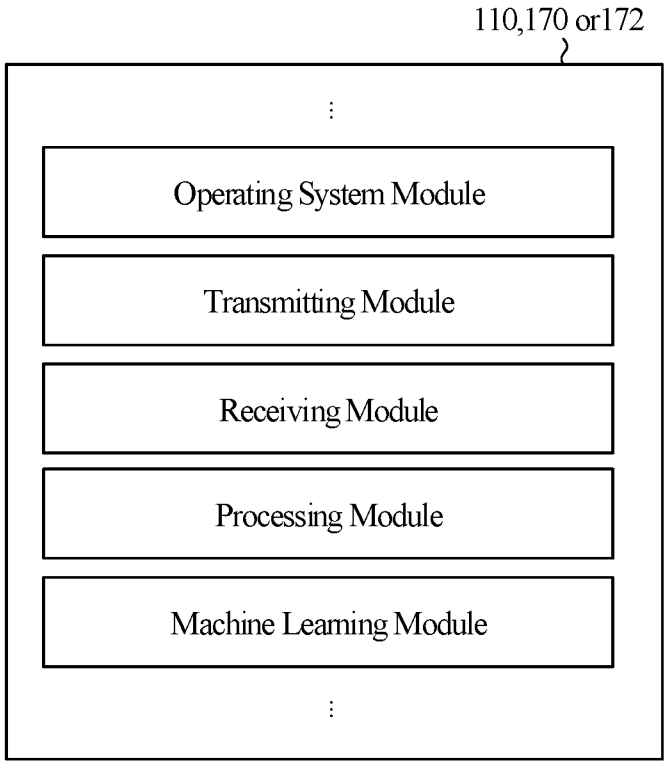

110,170 or172
Operating System Module
Transmitting Module
Receiving Module
Processing Module
Machine Learning Module

FIG.4

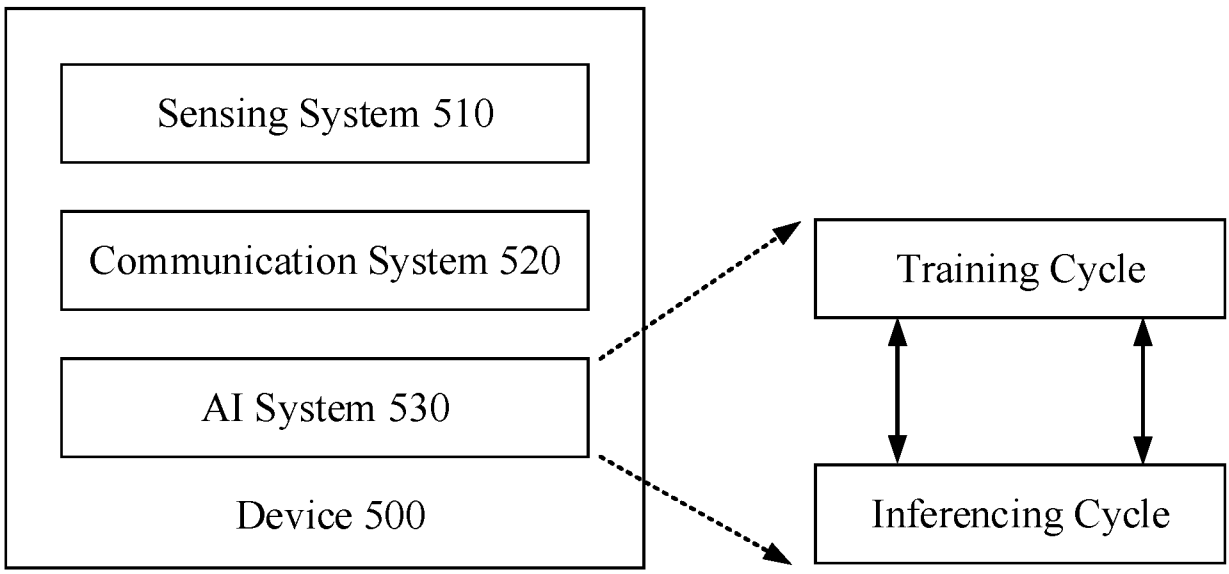

Sensing System 510
Communication System 520
AI System 530
Device 500
Training Cycle
Inferencing Cycle

FIG. 5

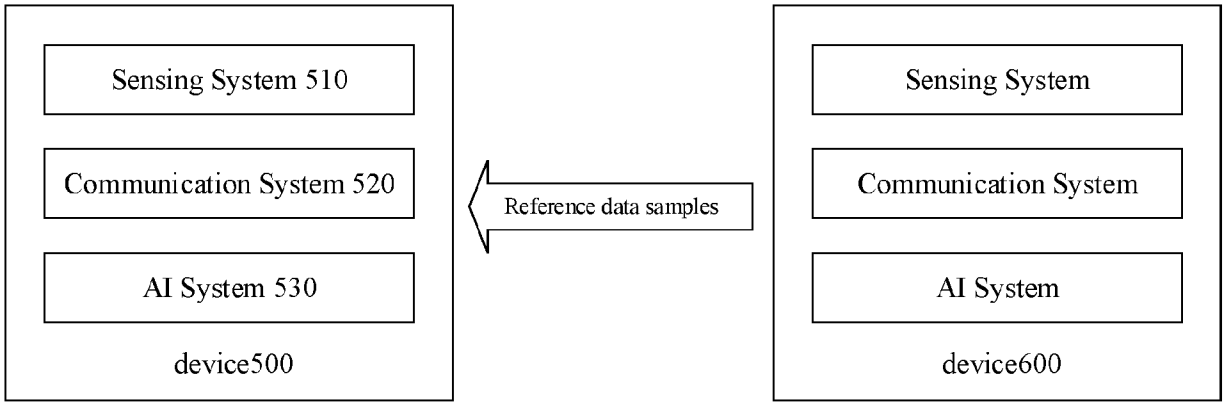

Sensing System 510
Communication System 520
AI System 530
device500
Reference data samples
Sensing System
Communication System
AI System
device600

FIG.6

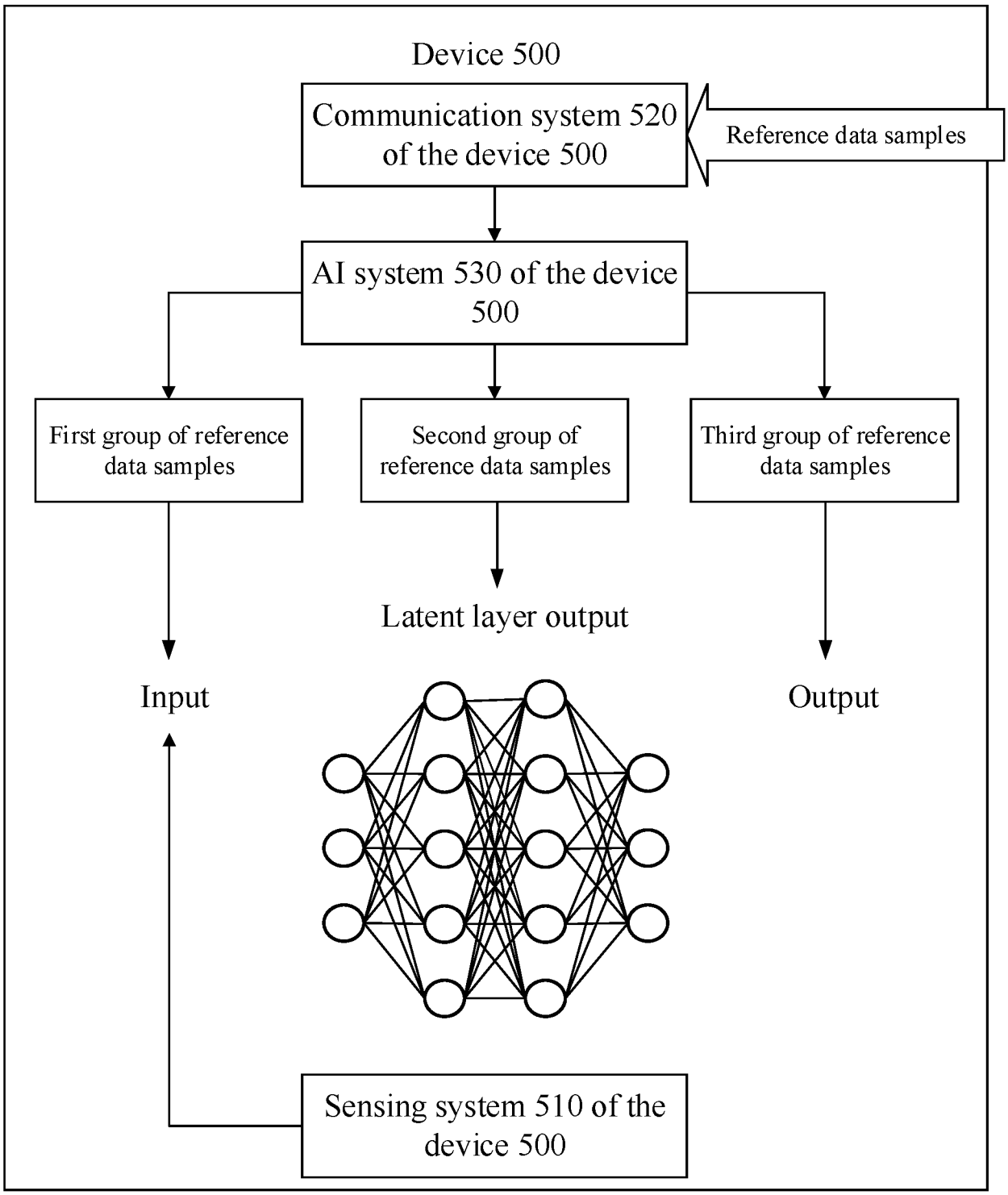

Device 500
Communication system 520 of the device 500
Reference data samples
AI system 530 of the device 500
First group of reference data samples
Second group of reference data samples
Third group of reference data samples
Latent layer output
Input
Output
Sensing system 510 of the device 500

FIG.7

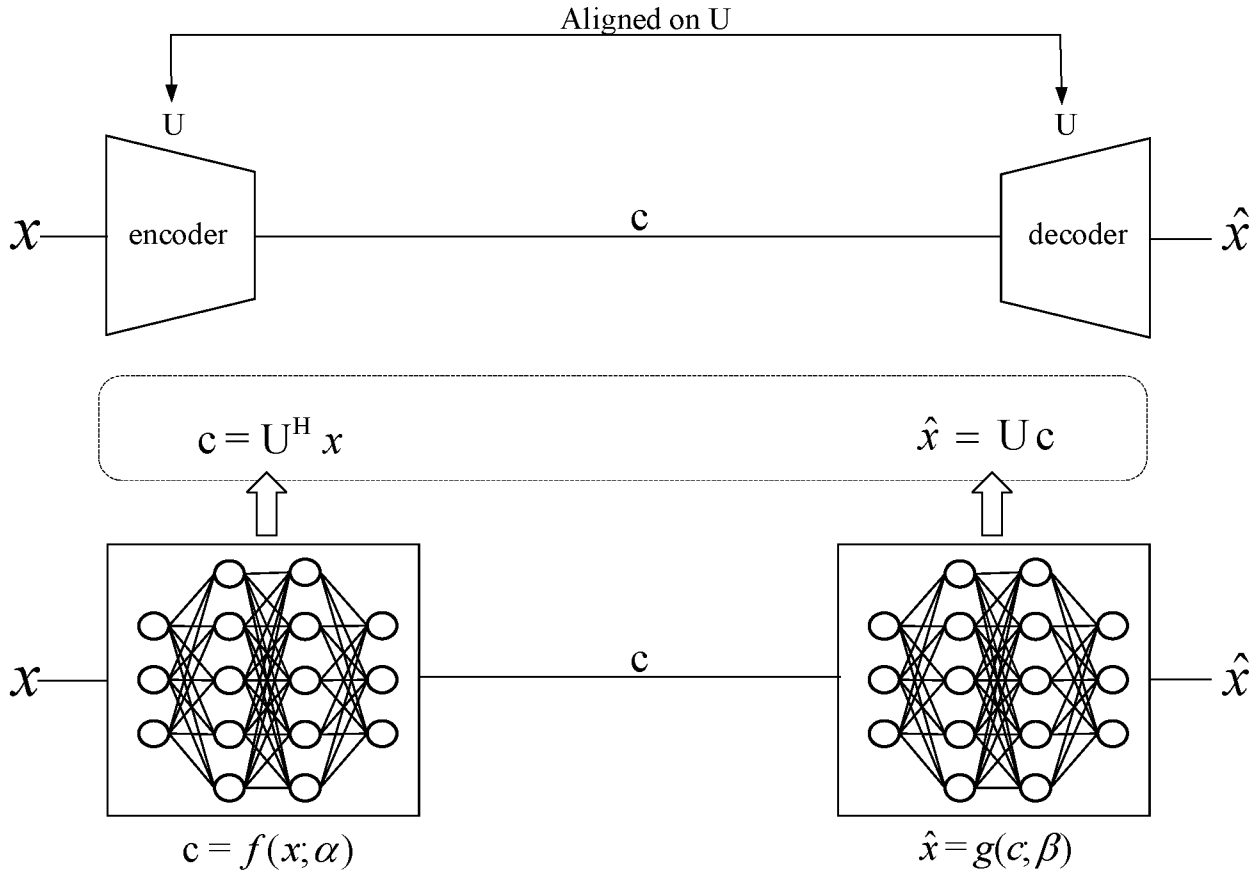
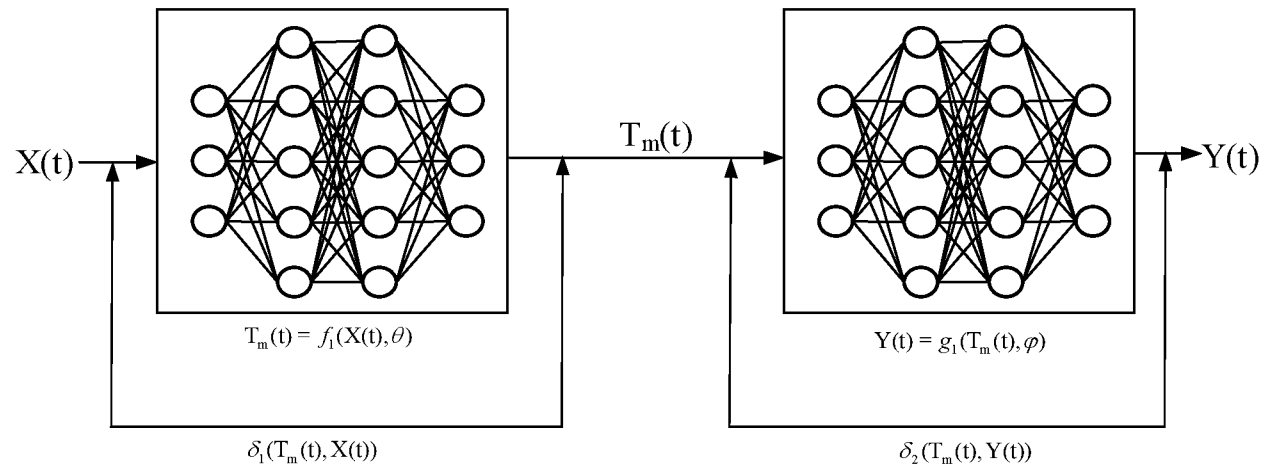


FIG.8



$$\rho_m(t) = \frac{\delta_1(T_m(t), X(t))}{\delta_2(T_m(t), Y(t))}$$

FIG.9

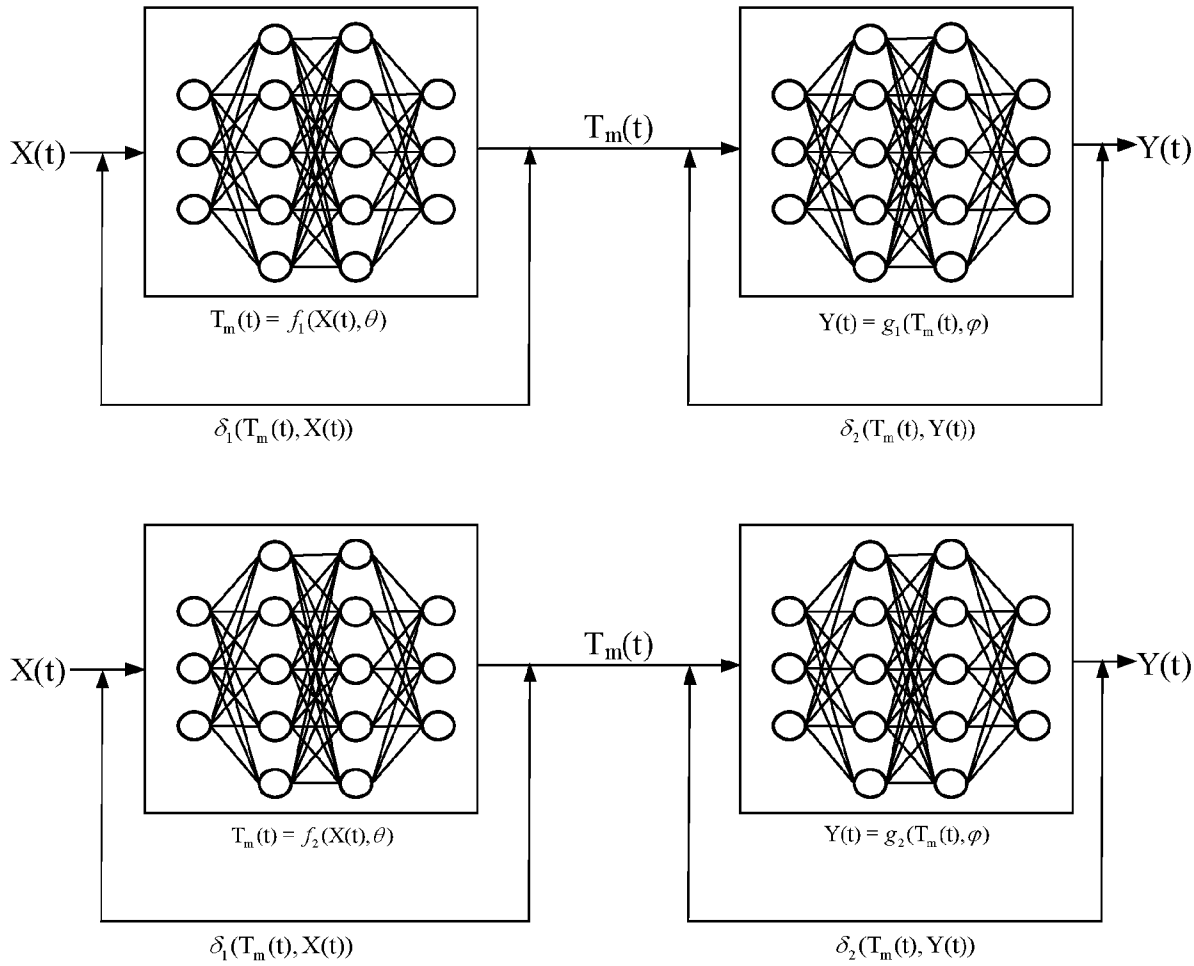

X(t)
$T_m(t)$
Y(t)
$T_m(t) = f_1(X(t), \theta)$
$Y(t) = g_1(T_m(t), \varphi)$
$\delta_1(T_m(t), X(t))$
$\delta_2(T_m(t), Y(t))$
X(t)
$T_m(t)$
Y(t)
$T_m(t) = f_2(X(t), \theta)$
$Y(t) = g_2(T_m(t), \varphi)$
$\delta_1(T_m(t), X(t))$
$\delta_2(T_m(t), Y(t))$

FIG.10

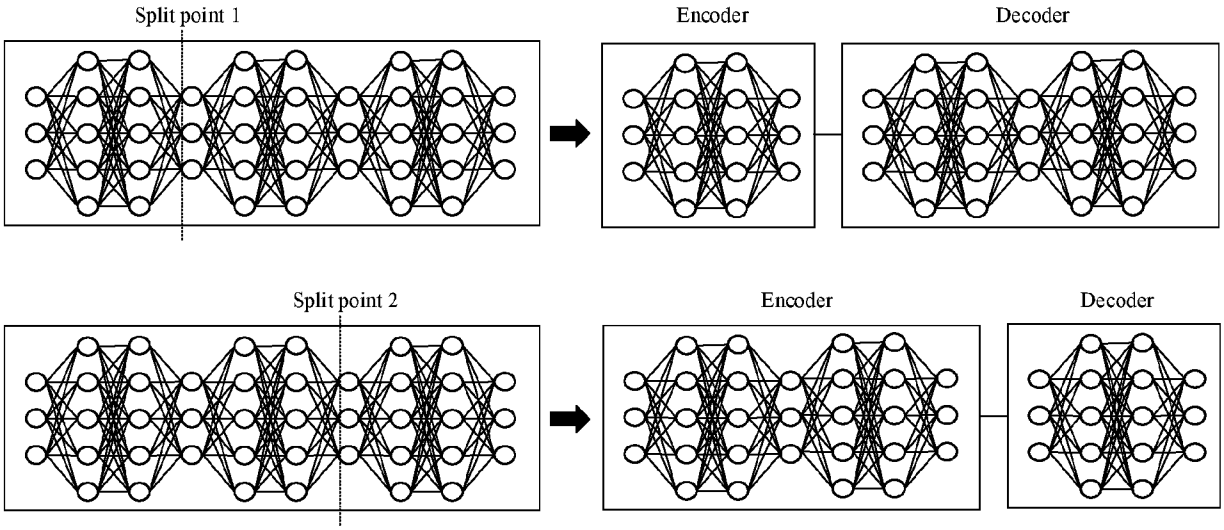

Split point 1
Encoder
Decoder
Split point 2
Encoder
Decoder

FIG.11

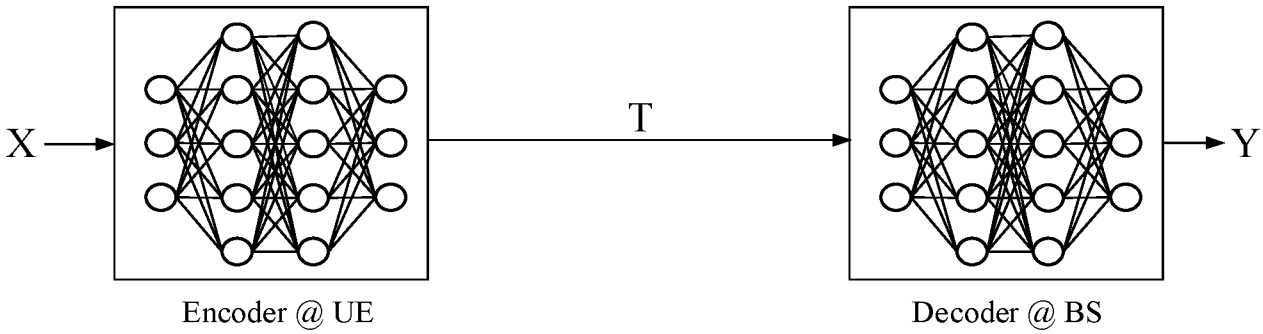


FIG. 12

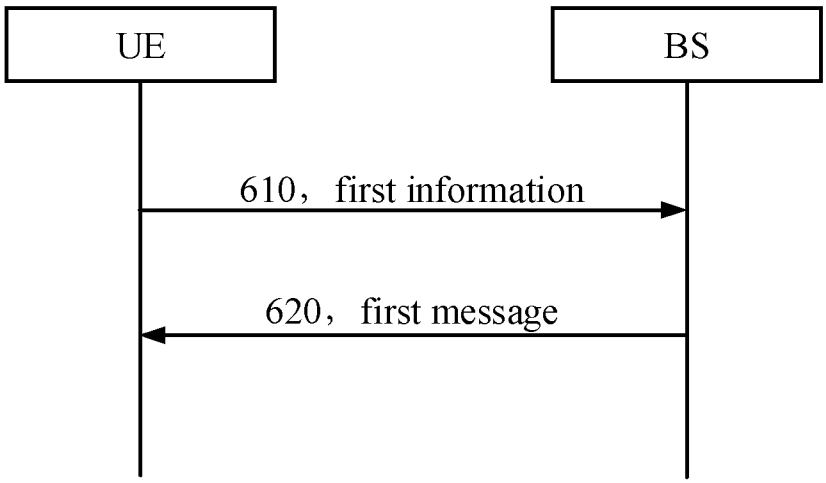


FIG.13

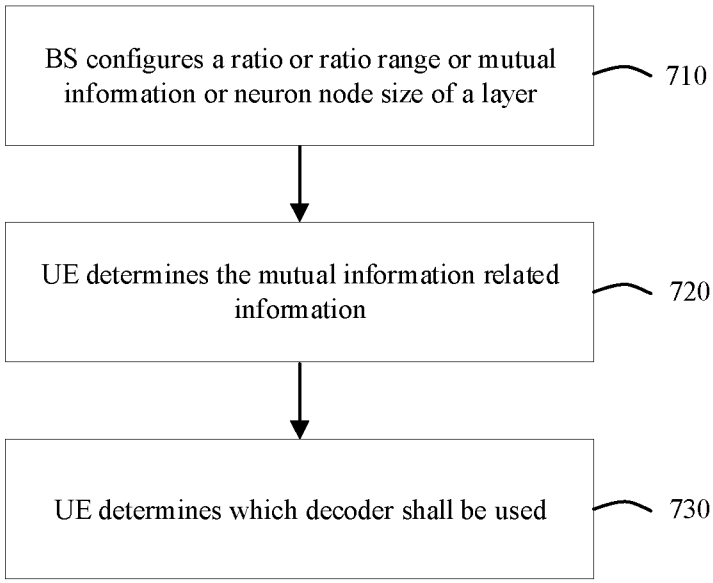
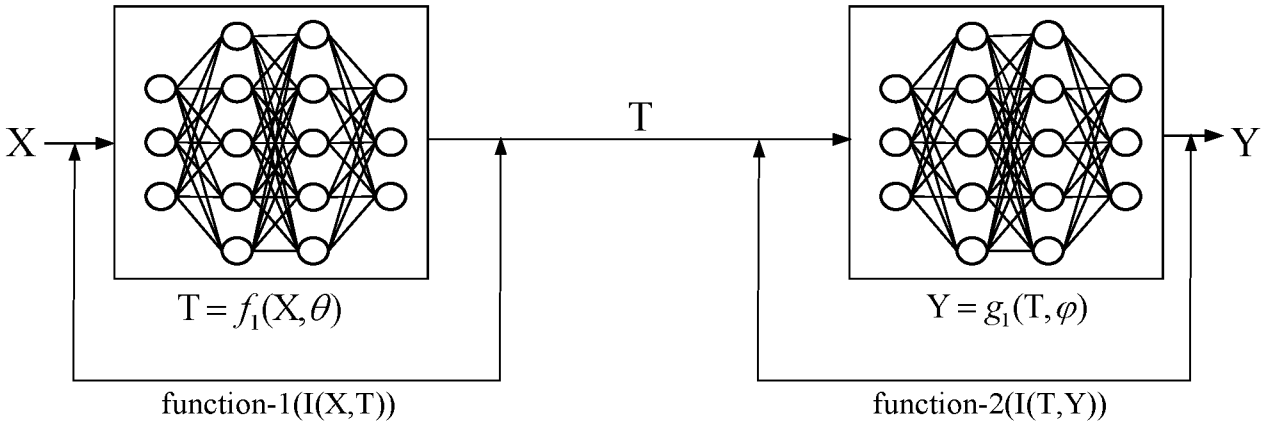


FIG.14



$$\alpha(T) = \frac{\text{function}-2\left(I\left(T,Y\right)\right)}{\text{function}-1\left(I\left(X,T\right)\right)}$$

FIG. 15

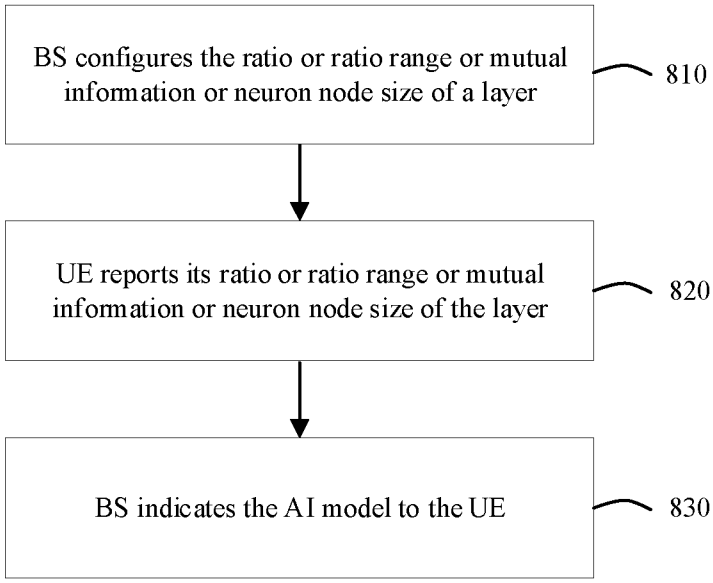


FIG.16

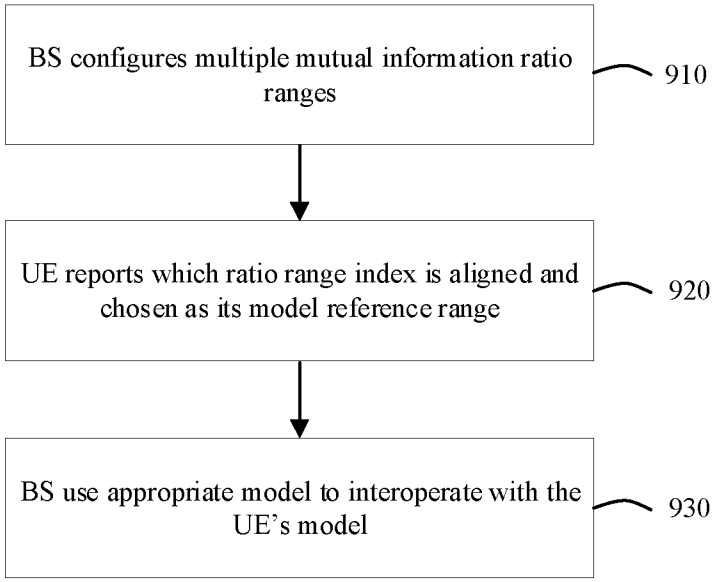


FIG.17

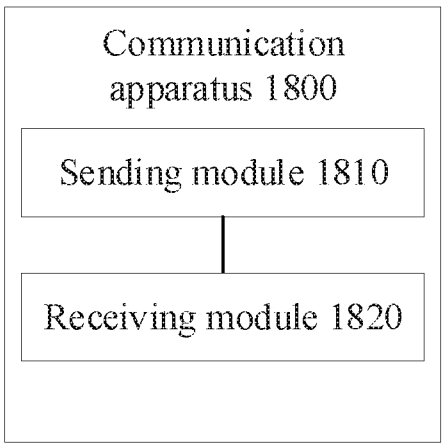


FIG. 18

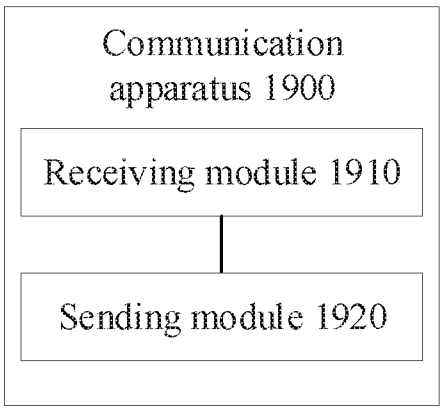


FIG. 19

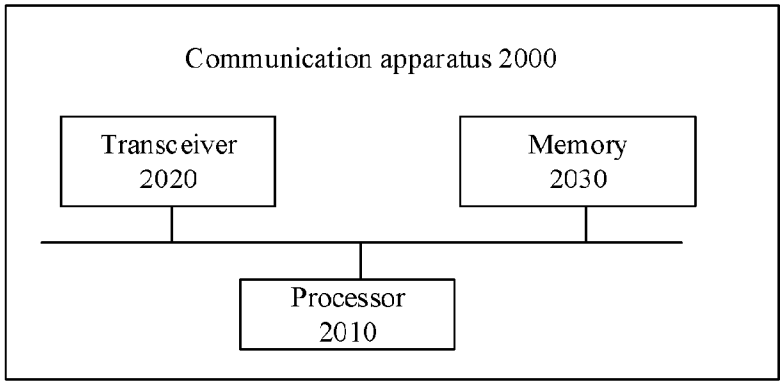

Communication apparatus 2000
Transceiver
2020
Memory
2030
Processor
2010

FIG. 20

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2023/124990

A. CLASSIFICATION OF SUBJECT MATTER

H04W24/02(2009.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC: H04L, H04W

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

ENTXT, CNTXT, DWPI, 3GPP: artificial intelligence, AI, model, mutual, indicate, range, index, neuron, ratio, HSIC, encoder, decoder

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 113747462 A (HUAWEI TECHNOLOGIES CO., LTD.) 03 December 2021 (2021-12-03) description, paragraphs [0020]-[0034], [0239]-[0271], figures 1-24	1-54
A	CN 115349279 A (BEIJING XIAOMI MOBILE SOFTWARE CO., LTD.) 15 November 2022 (2022-11-15) the whole document	1-54
A	US 2023045950 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 16 February 2023 (2023-02-16) the whole document	1-54
A	US 2023179490 A1 (HUAWEI TECHNOLOGIES CO., LTD.) 08 June 2023 (2023-06-08) the whole document	1-54
A	WO 2022133866 A1 (HUAWEI TECHNOLOGIES CO., LTD.) 30 June 2022 (2022-06-30) the whole document	1-54

☐ Further documents are listed in the continuation of Box C.

☑ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"D" document cited by the applicant in the international application

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
05 February 2024	**18 February 2024**

Name and mailing address of the ISA/CN	Authorized officer
CHINA NATIONAL INTELLECTUAL PROPERTY ADMINISTRATION **6, Xitucheng Rd., Jimen Bridge, Haidian District, Beijing 100088, China**	**WENG,YuQing** Telephone No. **(+86) 010-53961645**

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2023/124990

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	113747462	A	03 December 2021	US	2023087821	A1	23 March 2023
				WO	2021244334	A1	09 December 2021
				EP	4152797	A1	22 March 2023
CN	115349279	A	15 November 2022		None		
US	2023045950	A1	16 February 2023		None		
US	2023179490	A1	08 June 2023	EP	4181556	A1	17 May 2023
				WO	2022022334	A1	03 February 2022
				CN	114071484	A	18 February 2022
WO	2022133866	A1	30 June 2022	EP	4233303	A1	30 August 2023
				US	2023284139	A1	07 September 2023
				CN	116686278	A	01 September 2023