(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 2024/255040 A1

(51) International Patent Classification:
H04L 69/04 (2022.01)

(21) International Application Number:
PCT/CN2023/125044

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

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
63/507,872 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: **GE, Yiqun**; Huawei Administration Building, Bantian, Longgang, Shenzhen, Guangdong 518129 (CN). **TANG, Hao**; 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

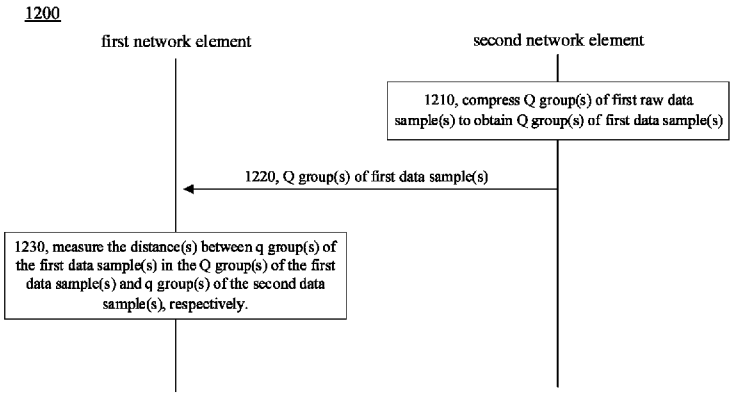

FIG. 12

(57) Abstract: Embodiments of the present application provide a communication method and a communication apparatus. The communication method includes: obtaining Q group (s) of first data sample (s) corresponding to Q layer (s) of an AI model, where the Q group (s) of the first data sample (s) is from compressed Q group (s) of first raw data sample (s) which is compressed according to Q transformation matrix (es), the Q group (s) of the first data sample (s) is related to an inference cycle of the AI model, and Q is a positive integer; and sending the Q group (s) of the first data sample (s). According to the above technical solution, the efficiency of data transmission can be improved.

WO 2024/255040 A1

COMMUNICATION METHOD AND COMMUNICATION APPARATUS

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] The present application is related to, and claims priority to, United States provisional patent application Serial No. 63/507,872, entitled "AI MODEL CROSS CONSISTENCE BY LATENT DATA REFERENCE CHECK ", 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 communications, and more specifically, to a communication method and a communication apparatus.

BACKGROUND

[0004] Artificial intelligence (AI)-based algorithms have been introduced into wireless communications to solve some wireless problems such as channel estimation, scheduling, channel state information (CSI) compression, positioning, beam-management, and so on. AI algorithm is a data-driven method that tunes some pre-defined architectures by a set of data samples called as training data set.

[0005] During the inference cycle of the AI model, data needs to be transmitted. Raw data may include user privacy. It may be against the privacy policy to transmit raw data. In addition, transmitting raw data may consume a lot of resources and is inefficient.

[0006] Therefore, an urgent technical problem that needs to be solved is how to improve data transmission efficiency.

SUMMARY

[0007] Embodiments of the present application provide a communication method and a communication apparatus. The technical solutions may improve data transmission efficiency.

[0008] According to a first aspect, an embodiment of the present application provides a communication method,

including obtaining Q group(s) of first data sample(s) corresponding to Q layer(s) of an AI model, where the Q group(s) of the first data sample(s) is from compressed Q group(s) of first raw data sample(s) which is compressed according to Q transformation matrix(es), the Q group(s) of the first data sample(s) is related to an inference cycle of the AI model, and Q is a positive integer; and sending the Q group(s) of the first data sample(s).

[0009] According to the above technical solution, the first data sample is a low-dimensional data sample which is compressed according to a transformation matrix. In this way, the bandwidth for the first data sample(s) can be saved and data transmission efficiency can be improved. At the same time, first raw data can be protected.

[0010] Each group may correspond to one layer of the AI model. Different groups may correspond to different layers.

[0011] In a possible design, the method further includes: sending first information indicating the Q transformation matrix(es).

[0012] Optionally, a transformation matrix be a unitary matrix or an orthonormal matrix.

[0013] Optionally, each basis vector of a transformation matrix may be a standard basis such as Fourier basis, DCT basis, wavelet basis, or the like.

[0014] In a possible design, the first information is further configured to indicate Q sampling matrix(es), the Q sampling matrix(es) is configured to sample Q group(s) of second raw data sample(s), and the Q transformation matrix(es) is configured to compress sampling result(s) of the Q group(s) of the second raw data sample(s) into Q group(s) of second data sample(s).

[0015] Optionally, a sampling matrix may be a random matrix or a pseudo-random matrix.

[0016] According to the above technical solution, the data sample can be obtained by compressing the raw data sample according to the sampling matrix and the transformation matrix. The dimensions of the sampling matrix and transformation matrix are smaller, which is beneficial to reducing the resources required for transmitting the sampling matrix and transformation matrix, thereby improving transmission efficiency.

[0017] In a possible design, the method further includes: receiving second information indicating difference(s) between q group(s) of second data sample(s) and q group(s) of the first data sample(s) in the Q group(s) of the first data sample(s), where the q group(s) of the second data sample(s) is based on inputs or outputs of q layer(s) in the Q layer(s) during the inference cycle, and q is a positive integer, $q \leqslant Q$.

[0018] For a first data sample and a second data sample corresponding to the same layer, the distance between the first data sample and the second data sample is approximately the same as the distance between the first raw data sample and the second raw data sample. In this way, computational complexity can be reduced, which is beneficial to improving processing efficiency.

[0019] In a possible design, the difference(s) between the q group(s) of the second data sample(s) and the q group(s) of

the first data sample(s) is configured to check whether the inference cycle is abnormal.

[0020] For example, if the distances corresponding to all the groups are consistently below the corresponding threshold(s), the current inference cycle may be considered normal.

[0021] According to the above technical solution, the difference(s) can be used to check whether the current inference cycle works as expected, which is conducive to ensuring the communication quality.

[0022] In addition, the inference cycle detection can be implemented with lower dimensional space. Compared to calculating the distance(s) between the first raw data sample(s) and the second raw data sample(s) in the original dimension, the dimensions of the first data sample(s) and second data sample(s) are lower, so the computational complexity can be reduced which is beneficial to improving processing efficiency.

[0023] In a possible design, the method further includes: sending third information indicating correspondence between the Q layer(s) and the Q group(s) of the first data sample(s).

[0024] In a possible design, the method further includes: sending fourth information indicating Q scoring function(s), where the Q scoring function(s) is configured to measure difference(s) between the Q group(s) of the first data sample(s) and Q group(s) of second data sample(s), and the Q group(s) of second data sample(s) is based on the inputs or outputs of the Q layer(s).

[0025] Optionally, each scoring function may be used to measure the distance between two samples.

[0026] Optionally, each scoring function may be used to measure the distance between two distributions.

[0027] According to a second aspect, an embodiment of the present application provides a communication method, including: receiving Q group(s) of first data sample(s) corresponding to Q layer(s) of an AI model, where the Q group(s) of the first data sample(s) is from compressed Q group(s) of first raw data sample(s) which is compressed according to Q transformation matrix(es), the Q group(s) of the first data sample(s) is related to an inference cycle of the AI model, and Q is a positive integer.

[0028] In a possible design, the method further includes: receiving first information indicating the Q transformation matrix(es).

[0029] In a possible design, the first information is further configured to indicate Q sampling matrix(es), the Q sampling matrix(es) is configured to sample Q group(s) of second raw data sample(s), and the Q transformation matrix(es) is configured to compress sampling result(s) of the Q group(s) of the second raw data sample(s) into Q group(s) of second data sample(s).

[0030] In a possible design, the method further includes: sending second information indicating difference(s) between q group(s) of second data sample(s) and q group(s) of the first data sample(s) in the Q group(s) of the first data sample(s), where the q group(s) of the second data sample(s) is based on inputs or outputs of q layer(s) in the Q layer(s) during the

inference cycle, and q is a positive integer, $q \leq Q$.

[0031] In a possible design, the difference(s) between the q group(s) of the second data sample(s) and the q group(s) of the first data sample(s) is configured to determine whether the inference cycle of the AI model is abnormal.

[0032] In a possible design, the method further includes: receiving third information indicating correspondence between the Q layer(s) and the Q group(s) of the first data sample(s).

[0033] In a possible design, the method further includes: receiving fourth information indicating Q scoring function(s), where the Q scoring function(s) is configured to measure difference(s) between the Q group(s) of the first data sample(s) and Q group(s) of second data sample(s), and the Q group(s) of the second data sample(s) is based on inputs or outputs of the Q layer(s).

[0034] According to a third aspect, a communication apparatus is provided. The communication apparatus includes a function or unit configured to perform the method according to the first aspect or any one of the possible designs of the first aspect.

[0035] For example, the communication apparatus may be a network device or a chip in the network device. For another example, the communication apparatus may be a terminal device or a chip in the terminal device.

[0036] According to a fourth aspect, a communication apparatus is provided. The communication apparatus includes a function or unit configured to perform the method according to the second aspect or any one of the possible designs of the second aspect.

[0037] For example, the communication apparatus may be a terminal device or a chip in the terminal device. For another example, the communication apparatus may be a network device or a chip in the network device.

[0038] According to a fifth aspect, a system is provided. The system includes: the communication apparatus according to the third aspect and the communication apparatus according to the fourth aspect.

[0039] According to a sixth aspect, a communication apparatus is provided. The communication apparatus includes at least one processor, and the at least one processor is coupled to at least one memory. The at least one memory is configured to store a computer program or one or more instructions. The at least one processor is configured to: invoke the computer program or the one or more instructions from the at least one memory and run the computer program or the one or more instructions, so that the communication apparatus performs the method in any one of the first aspect or the possible designs of the first aspect, or the communication apparatus performs the method in any one of the second aspect or the possible designs of the second aspect.

[0040] For example, the communication apparatus may be a network device or a component (for example, a chip or integrated circuit) installed in the network device. For another example, the communication apparatus may be a terminal device

or a component (for example, a chip or integrated circuit) installed in the terminal device.

[0041] According to a seventh aspect, a communication apparatus is provided. The communication apparatus includes a processor and a communications interface. The processor is connected to the communications interface. The processor is configured to execute the one or more instructions, and the communications interface is configured to communicate with other network elements under the control of the processor. The processor is enabled to perform the method according to the first aspect or any one of the possible designs of the first aspect, or the second aspect or any one of the possible designs of the second aspect.

[0042] According to an eighth aspect, a computer storage medium is provided. The computer storage medium stores program code, and the program code is used to execute one or more instructions for the method according to the first aspect or any one of the possible designs of the first aspect, or the second aspect or any one of the possible designs of the second aspect.

[0043] According to a ninth aspect, the present application provides a computer program product including one or more instructions, where when the computer program product runs on a computer, the computer performs the method according to the first aspect or any one of the possible designs of the first aspect, or the second aspect or any one of the possible designs of the second aspect.

DESCRIPTION OF DRAWINGS

[0044] FIG. 1 is a schematic diagram of an application scenario according to the present application;

[0045] FIG. 2 illustrates an example communication system 100;

[0046] FIG. 3 illustrates an example device in the communication system;

[0047] FIG. 4 is a schematic diagram of a device in two cycles according to an embodiment of the present application;

[0048] FIG. 5 illustrates example local data of a device according to an embodiment of the present application;

[0049] FIG. 6 is a schematic diagram of the working situation of an AI model;

[0050] FIG. 7 is a schematic diagram of an example scenario;

[0051] FIG. 8 illustrates an example data transmission between two devices according to an embodiment of the present application;

[0052] FIG. 9 is a schematic diagram of three groups of reference data sample(s) according to an embodiment of the present application;

[0053] FIG. 10 is a schematic diagram of an example distance calculation according to an embodiment of the present application;

[0054] FIG. 11 is schematic diagram of two examples of encoders according to an embodiment of the present application;

[0055] FIG. 12 is a schematic flowchart of a communication method according to an embodiment of the present application;

[0056] FIG. 13 is a schematic diagram of an example compression process of a reference data sample according to an embodiment of the present application;

[0057] FIG. 14 is a schematic diagram of an example X according to an embodiment of the present application;

[0058] FIG. 15 is a schematic diagram of an example compression process according to an embodiment of the present application;

[0059] FIG. 16 is a schematic diagram of an example distance on the low spectrum space according to an embodiment of the present application;

[0060] FIG. 17 is a schematic diagram of the autoencoder with one group of reference data samples according to an embodiment of the present application;

[0061] FIG. 18 is a schematic diagram of three groups of reference data samples according to an embodiment of the present application; and

[0062] FIGS. 19-23 are schematic block diagrams of possible devices according to embodiments of the present application.

DESCRIPTION OF EMBODIMENTS

[0063] The following describes technical solutions of the present application with reference to the accompanying drawings.

[0064] The embodiments of the present invention may be applied to communication systems of next generation (e.g. sixth generation (6G) or later), 5th Generation (5G), new radio (NR), long term evolution (LTE), or the like.

[0065] FIG. 1 is a schematic structural diagram of an example communication system.

[0066] Referring to FIG.1, as an illustrative example without limitation, a simplified schematic illustration of a communication system is provided. A communication system 100 includes a radio access network 120. The radio access network 120 may be a next generation (e.g. 6G or later) radio access network, or a legacy (e.g. 5G, 4G, 3G or 2G) radio access network. One or more communication electric device (ED) 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.

[0067] FIG. 2 is a schematic structural diagram of another example communication system.

[0068] In general, a 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.

[0069] The terrestrial communication system and the non-terrestrial communication system could be considered sub-systems of the communication system. In the example shown, the communication system 100 includes electronic devices (ED) 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.

[0070] 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.

[0071] 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.

[0072] 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.

[0073] 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.

[0074] 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.

[0075] 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 personal communications service (PCS) phone, a session initiation protocol phone, a wireless local loop (WLL) station, 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. 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.

[0076] 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.

[0077] 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.

[0078] 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.

[0079] Artificial intelligence (AI) technologies can be applied in communication, including artificial intelligence or machine learning (AI/ML) based communication in the physical layer and/or AI/ML based communication in the higher layer, such as medium access control (MAC) layer. For example, in the physical layer, the AI/ML based communication may aim to

optimize component design and/or improve the algorithm performance. For example, AI/ML may be applied in relation to the implementation of channel coding, channel modelling, channel estimation, channel decoding, modulation, demodulation, multiple-input multiple-output (MIMO), waveform, multiple access, physical layer element parameter optimization and update, beam forming, tracking, sensing, and/or positioning, etc. For the MAC layer, the AI/ML based communication may aim to utilize the AI/ML capability for learning, prediction, and/or making decisions to solve a complicated optimization problem with possible better strategy and/or optimal solution, e.g. to optimize the functionality in the MAC layer. For example, AI/ML may be applied to implement: intelligent transmission and reception point (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.

[0080] In order to facilitate understanding of the embodiments of the present application, terms related to AI/ML that may be involved in the embodiments of the present application are described below.

[0081] (1) Data collection

[0082] Data is a very important component for AI/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.

[0083] (2) AI/ML model training

[0084] 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.

[0085] (3) AI/ML model inference

[0086] A process of using a trained AI/ML model to produce a set of outputs based on a set of inputs.

[0087] (4) AI/ML model validation

[0088] 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 selecting model parameters that generalize beyond the dataset used for model training. The model parameter after training can be adjusted further by the validation process.

[0089] (5) AI/ML model testing

[0090] 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.

[0091] (6) Online training

[0092] 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.

[0093] (7) Offline training:

[0094] 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.

[0095] (8) AI/ML model delivery/transfer

[0096] AI/ML model delivery/transfer is a generic term referring to 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.

[0097] (9) Life cycle management (LCM)

[0098] 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 the sustainable operation of AI/ML in the NR air-interface.

[0099] 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.

[0100] 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.

[0101] (10) Supervised learning

[0102] 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.

[0103] (11) Federated learning (FL)

[0104] Federated learning 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"). The central node can also be called the central device. The edge nodes can also be called worker or worker devices. The central device is connected to the worker devices.

[0105] According to the wireless FL technique, a central node 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 central node, a set of AI/ML model parameters that describe the local AI/ML model.

[0106] 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 central node may aggregate the local AI/ML model parameters reported from the plurality of edge nodes and, based on such aggregation, update the global AI/ML model. A subsequent iteration progresses much like the first iteration. The central node 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, for example, the AI/ML model is converged or the training stopping conditions are satisfied.

[0107] The wireless FL technique does not involve the exchange of local data samples. Indeed, the local data samples remain at respective edge nodes.

[0108] AI-based algorithms have been introduced into wireless communications to solve a number of wireless problems such as channel estimation, scheduling, CSI compression (from UE to BS), beamforming for MIMO, localization, and so on. AI algorithms are a data-driven approach to tuning some predefined architectures by a set of data samples called training data sets.

[0109] Neural networks are a typical way to implement AI algorithms. Deep neural network (DNN) is taken as an example, the DNN can be trained with the training data sets to obtain a model for inference. Recent AI trains DNN architectures by setting up neurons with stochastic gradient descent (SGD) algorithms. For example, DNN includes CNN, RNN, transformers, and the like.

[0110] A communication system includes a plurality of connected devices. For example, a device may be a BS or UE. For example, the communication system may be the communication system 100 in FIG. 1 or FIG. 2, and the devices can be the network elements shown in FIG. 1 or FIG. 2.

[0111] FIG. 3 is a schematic structural diagram of a device according to an embodiment of the present application. As shown in FIG. 3, the device may include at least one of sensing module, communication module, or AI module. The sensing module may be configured to sense and collect signals and/or data. The communication module may be configured to transmit

and receive signals and/or data. The AI module may be configured to train and/or reason the AI implementations.

[0112] In order to facilitate understanding of the embodiment of the present application, DNN is taken as an example to illustrate an AI implementation in an embodiment of the present application.

[0113] An exemplary AI implementation is DNN-based in two cycles: a training cycle and an inference cycle. The training cycle may also be called the learning cycle. The inference cycle may also be called the reasoning circle.

[0114] FIG. 4 is a schematic diagram of a device in two cycles according to an embodiment of the present application.

[0115] As an example, during an inference cycle, the AI module of the device may perform one inference or a series of inferences with one or more DNNs to fulfill one or more tasks, where the sensing module of the device may generate signals and/or data and the communication module of the device may receive the signals and/or data from other device or devices. For example, the inputs of the one or more DNNs may be the signals and/or data generated by the sensing module of the device, and/or the signals and/or data received by the communication module of the device. After the AI module of the device finishes inferencing, the communication module of the device may transmit the inferencing results to other device or devices.

[0116] As another example, during a training cycle, the AI module of the device may train one or more DNNs, where the sensing module of the device may generate signals and/or data and the communication module of the device may receive the signals and/or data from other device or devices. For example, the training data of the one or more DNNs may be the signals and/or data generated by the sensing module of the device, and/or the signals and/or data received by the communication module of the device. During and/or after the AI module finishes training, the communication module of the device may transmit the training results to other device or devices.

[0117] The AI implementations may either switch between the two cycles or stay in the two cycles simultaneously.

[0118] For example, the AI module of the device may train a DNN during the training cycle. And at the end of the training cycle, the AI implementation switches to the inference cycle, which means the AI module performs inference on that trained DNN. At the end of the inference cycle the AI implementation switches to the training cycle again, and so on.

[0119] For another example, the AI module of the device may train a second DNN but still perform inference on a first DNN.

[0120] The device mentioned above is merely an example, and the way in which the modules are divided and the number of modules in FIG. 3 and FIG. 4 do not constitute any limitation to the embodiments of the present application. For example, a communication module may be replaced by two modules, i.e., a transmitting module and a receiving module. The transmitting module may be configured to transmit signals and/or data, and the receiving module may be configured to receive signals and/or data. For another example, the sensing module and the communication module may be integrated as one module. For another example, the device may also include a processing module. The processing module may be configured to process

signals and/or data. For another example, the device may not include the AI module. For another example, the AI module may only be configured to reason the AI implementation, or the AI module only stays in the inference cycle.

[0121] Wireless systems may support AI in both learning and inferencing cycles for generalization and interconnections.

[0122] FIG. 5 shows example local data of a device. The local data of a device may include at least one of the following: local sensing data provided by the sensing module of the device, local channel data provided by the communication module of the device, local AI model data provided by the AI module of the device, or local latent output data provided by the AI module of the device. The local channel data is based on the measurement results of the channel. The local channel data can also be considered as sensing results. Thus, the local channel data can be considered as provided by the communication modules or sensing module.

[0123] For example, as shown in FIG. 5, the local sensing data may include at least one of RGB data, Lidar data, temperature, air pressure, or electric outrage.

[0124] For example, as shown in FIG. 5, the local channel data may include at least one of channel state information (CSI), received signal strength indication (RSSI), or delay.

[0125] The local AI model data can also be referred to as neuron data. For example, as shown in FIG. 5, the local AI model data may include at least one of the following: part or all of the neurons in the local AI model(s) deployed on the device or part or all of gradients of the local AI model(s) deployed on the device. Neurons can be considered as functions including weights.

[0126] For example, as shown in FIG. 5, the local latent output data may include one or more latent outputs of the local AI model(s) deployed on the device.

[0127] A device may receive the local data of one or more other devices. As an example, the data received by the communication module of the device may include at least one of sensing data of one or more other devices, channel data of one or more other devices, AI model data of one or more other devices, or latent output data of one or more other devices.

[0128] For example, the data received by the communication module of device #A may include channel data of device #B and device #C, and AI model data of device #C. The channel data of device #B and device #C refer to the local channel data of device #B and the local channel data of device #C. The AI model data of device #C refers to the local AI model data of device #C. Device #A, device #B, and device #C are different devices.

[0129] For example, sensing data received by the communication module may include at least one of RGB data, Lidar data, temperature, air pressure, or electric outrage.

[0130] For example, channel data received by the communication module may include at least one of CSI, RSSI, or delay.

[0131] For example, AI model data received by the communication module may include at least one of part or all of the neurons in the AI model(s), or part or all of gradients of the AI model(s).

[0132] For example, latent output data received by the communication module may include one or more latent outputs of the AI model(s).

[0133] Whether the AI model deployed on a device can work is crucial for communication quality.

[0134] As a data-driven method, an AI model inevitably suffers from low generalization. If a real-world sample, such as a user data sample, is outlier to the training data set, the AI model wouldn't make a good inference on the real-world sample. Moreover, even given an outlier input, the AI model may not detect it.

[0135] For example, in wireless communication, the user device is moving. The AI model deployed on the user device may work in some environments, but may not work in others, which can affect the communication quality.

[0136] FIG. 6 is a schematic diagram of the working situation of an AI model.

[0137] As shown in FIG. 6, when the user data sample collected by the user device is within the zone of the training samples used to train the AI model, the AI model can work. As the user device moves, the user data sample collected by the user device may be outside the zone of the training samples, and the AI model doesn't work.

[0138] In wireless communication, AI models deployed on different devices may need to work together. Dual sided model is taken as an example. Dual sided model may be in a form of AE, whose encoding DNN is on transmitter side and decoding DNN on receiver side. The encoding DNN and decoding DNN are likely trained and provided by different providers. Moreover, it is hard for AI providers to open their DNN models. This may result in the AI models not working together.

[0139] FIG. 7 is a schematic diagram of an example scenario.

[0140] As shown in FIG. 7, an encoder deployed on UE and a decoder deployed on BS need to work together. However, the encoder and the decoder may be trained independently by different providers, e.g. provider #1 and provider #2 in FIG. 7, which may affect their interconnection.

[0141] The embodiment of the present application provides a communication method that ensures that the AI model can work through the comparison between reference data and local data, thereby improving the communication performance. The reference data can also refer to a reference signal. The local data can also refer to a local signal. For the convenience of description, no distinction will be made in the embodiments of the present application.

[0142] During the inference cycle, the AI module of a device may work in a single user mode or cooperative mode. In both modes, the device may receive reference data sample(s) from one or more other devices. Or the reference data sample(s) may be pre-stored on the device.

[0143] The type of the local data sample(s) may be related to any type of the data mentioned in FIG. 5. For example,

the local data sample(s) may be corresponding to Lidar data. For another example, the local data sample(s) may be corresponding to CSI.

[0144] The local data sample(s) generated by one device can be transmitted to another device as reference data sample(s) for the AI model on another device.

[0145] For example, device #1 may receive reference data sample(s) from device #2. The local data sample(s) generated by the device #2 can be regarded as the reference data sample(s) for the AI model on device #1.

[0146] The reference data sample(s) may be related to any type of the data received by the communication module of the device mentioned above. For example, the reference data sample(s) may be corresponding to Lidar data. For another example, the reference data sample(s) may be corresponding to CSI.

[0147] In the case of receiving a plurality of groups of reference data sample(s), the type of the data may be the same.

[0148] FIG. 8 shows an example of the data transmission between two devices.

[0149] Specifically, a device may receive Q group(s) of reference data sample(s) from another device. Q is a positive integer.

[0150] In the case of receiving a plurality of groups of reference data sample(s), the number of reference data samples in each group can be the same or different.

[0151] For example, other device(s) may transmit Q group(s) of reference data sample(s) in broadcast, multicast, or unicast channels.

[0152] The Q group(s) of reference data sample(s) corresponds to Q group(s) of local data sample(s), respectively. The distance between each group in the Q group(s) of reference data sample(s) and the corresponding group in the Q group(s) of local data sample(s) may be measured.

[0153] The Q group(s) of reference data sample(s) may correspond to Q layer(s) of AI model(s), respectively. One group of reference data sample(s) corresponds to one layer, which may be understood as the group of reference data sample(s) corresponds to the inputs or outputs of the layer. Correspondingly, the Q group(s) of local data sample(s) may be based on the Q layer(s) of AI model(s). For each group of the reference data samples(s), the corresponding group of local data sample(s) is based on the layer corresponding to the group of the reference data sample(s). The local data sample(s) may be sampled from the local data related to the layer(s). The local data may be the inputs or outputs of the Q layer(s). The Q group(s) of local data sample(s) may be sampled from the inputs or outputs of the Q layer(s). For example, one group of reference data sample(s) corresponds to the inputs of an AI model, in which case, the corresponding group of local data sample(s) may be obtained by sampling the inputs of the AI model.

[0154] As an example, the AI module of the device may randomly, non-randomly, uniformly, or non-uniformly sample

its local data related to the Q layer(s) to obtain the Q group(s) of local data sample(s).

[0155] The Q group(s) of reference data sample(s) may be related to Q layer(s) of one or more AI models. For the convenience of description, in the embodiments of present application, only the Q layers belonging to one AI model are used as an example for explanation.

[0156] FIG. 9 is a schematic diagram of three groups of reference data sample(s).

[0157] For example, as shown in FIG.9, there are three groups of reference data sample(s) received by the communication module of the device #1. The three groups of reference data sample(s) may be processed by the AI module of the device #1. The first group corresponds to the input layer of an AI model, the second group corresponds to one latent layer of the AI model, and the third group corresponds to the output layer of the AI model. Specifically, the first group corresponds to the inputs of the AI model, the second group corresponds to one latent layer outputs of the AI model, and the third group corresponds to the outputs of the AI model. The AI model may be a local AI model of the device #1. The first group of local data sample(s) may be sampled from the inputs to the AI model, the second group of local data sample(s) may be sampled from the latent layer outputs and the third group of local data sample(s) may be sampled from the outputs from the AI model. For example, as shown in FIG. 9, the inputs of the AI model may include the local sensing data provided by the sensing module of the device #1.

[0158] FIG. 9 is merely an example and shall not constitute any limitation on the present application. For example, the inputs of the AI model may also include data from other sources, such as data received by the communication module of the device #1. For another example, the inputs of the AI model may include the data that has been preprocessed for the local sensing data provided by the sensing module of the device #1. For another example, the number of groups of reference data sample(s) may be other values. The three groups of reference data sample(s) may be related to other layers.

[0159] The reference data sample(s) may be used to determine whether the current inference procedure is abnormal or not. In other words, the reference data sample(s) may be used to determine whether the current inference procedure is working as expected.

[0160] The following describes examples of application scenarios for the reference data sample(s).

[0161] If the AI model does not work during the inference cycle of as expected, it may be damaged, it may not be suitable for the current data, for example, the AI model may be outdated, or it may not be able to work with other AI models. The abnormal inference cycle of the AI model may lead to incorrect inference results, which may affect the relevant data processing results or data transmission quality.

[0162] The distance(s) between the local data sample(s) and the reference data sample(s) can be used to check whether the current inference cycle works as expected, which is conducive to ensuring the communication quality.

[0163] In some scenarios, as the device moves, the local data collected by the device may be outside the zone of the training samples, statistically outliers, and the AI model deployed on the device doesn't work.

[0164] The distance(s) between the local data sample(s) and the reference data sample(s) can be used to check whether the AI model can work. In other words, the distance(s) between the local data sample(s) and the reference data sample(s) can be used to check generalization of the AI model.

[0165] As an example, the reference data sample(s) may be related to the training data of the AI model.

[0166] Exemplarily, AI model # A can be a trained model. The reference data sample(s) may be generated when the AI model # A performs inference on target data. The target data is within the training data range, so the likelihood of the AI model #A's inference process working properly is higher. Based on the inputs, outputs, and/or latent layer outputs of the AI model during this inference process, the reference data sample(s) can be generated. The closer the local data sample(s) of an AI model is to the reference data sample(s), the greater the likelihood that the AI model can work.

[0167] Reference data sample(s) can also be determined through other methods. The embodiments of the present application do not limit this.

[0168] In the embodiments of the present application, the distance between the reference distribution and the distribution of the latent layer can be used to check whether AI model can work with the current local data, which is conducive to ensuring the quality of data processing or communication.

[0169] In some scenarios, a plurality of AI models need to work together. For example, the output of a latent layer of one AI model may be the input of a latent layer of another AI model. These AI models may be trained independently by different providers.

[0170] The distance(s) between the local data sample(s) and the reference data sample(s) can be used to check whether a plurality of AI models that need to work together can work together. In other words, the distance(s) between the local data sample(s) and the reference data sample(s) can be used to check the interconnection or cross consistency of the AI models.

[0171] The closer the local data sample(s) of an AI model is to the local data sample(s) of another AI model, the greater the likelihood that the two AI models can work together.

[0172] For example, for two AI models with the same structure (such as AE #A and AE #B), the smaller the distance between the local data sample(s) corresponding to the output of the encoder of the AE #A and the local data sample(s) corresponding to the output of the encoder of the AE #A, the higher the possibility that the two AI models can work together, that is, the output of the encoder of the AE #A can be used as the input of the decoder of AE #B, or, the output of the encoder of the AE #B can be used as the input of the decoder of AE #A.

[0173] In some embodiments, for two AI models that need to work together, the reference data sample(s) may be

sampled from the outputs of latent layer in one of the AI models. The distance between the reference data sample(s) and the local data sample(s) corresponding to the latent layer of another AI model can be used to check interconnection.

[0174] For example, for two AI models with the same structure (such as AE #A and AE #B), the reference data sample(s) may be sampled from the output of the encoder of AE #A, and the local data sample(s) may be sampled from the output of the encoder of AE #B. In this case, the smaller the distance between the reference data sample(s) and the local data sample(s), the greater the likelihood that the two AI models can work together.

[0175] Reference data sample(s) can also be determined through other methods. The embodiments of the present application do not limit this.

[0176] In the embodiments of the present application, the distance between the reference distribution and the distribution of the latent layer can be used to check whether AI models can work together, which is conducive to ensuring the quality of data processing or communication.

[0177] The Q layer(s) may belong to one or more local AI models deployed on the device. The embodiments of the present application do not limit the number of local AI models. For the convenience of description, the embodiments of the present application mainly use a local AI model as an example for explanation, and the implementation methods of other local AI models can refer to this local AI model.

[0178] Specifically, the distance(s) between the Q group(s) of reference data sample(s) and the corresponding group(s) of local data sample(s) may be used to determine whether the AI model works as expected.

[0179] Optionally, the device may measure the distance(s) between the local data sample(s) and the reference data sample(s) group by group to obtain Q distance(s) corresponding to the Q group(s). And then the Q distance(s) may be used to determine whether the AI model works as expected.

[0180] Alternatively, the device may measure the distance(s) between the local data sample(s) and the reference data sample(s) group by group to obtain q distance(s) corresponding to q group(s) in the Q group(s). In other words, the device may calculate distance based on a portion of the Q group(s). And then the q distance(s) may be used to determine whether the AI model works as expected.

[0181] The relationship between the distance(s) and the inference cycle can be set as needed.

[0182] For example, the greater the distance(s), the greater the likelihood of the inference cycle being abnormal. For the convenience of description, the embodiments of the present application will only be explained using this as an example.

[0183] The conditions for determining whether the AI model works as expected can be set as needed.

[0184] For example, if the distance(s) corresponding to all the group(s) is consistently below the corresponding threshold(s), the current inference procedure may be considered normal. Otherwise, the current inference procedure may be

considered abnormal. In the case of a plurality of groups of reference data sample(s), the thresholds corresponding to different groups can be the same or different. The threshold(s) may be pre-defined. Or the threshold(s) may be received by the device. Or the threshold(s) may be determined by the device itself.

[0185] For another example, if the distance(s) corresponding to all the group(s) is consistently greater than or equal to the corresponding threshold(s), the current inference procedure may be considered abnormal. Otherwise, the current inference procedure may be considered normal. In the case of a plurality of groups of reference data sample(s), the thresholds corresponding to different groups can be the same or different. The threshold(s) may be pre-defined. Or the threshold(s) may be received by the device. Or the threshold(s) may be determined by the device itself.

[0186] For another example, in the case of a plurality of groups of reference data sample(s), if the average distance of all the groups is below a threshold, the current inference procedure may be considered normal. Otherwise, the current inference procedure may be considered abnormal. The threshold may be pre-defined. Or the threshold may be received by the device from the other device. Or the threshold may be determined by the device itself.

[0187] The above conditions are merely examples. Other conditions about the above distance can be set to determine whether the inference procedure works as expected.

[0188] FIG. 10 is a schematic diagram of an example distance calculation. The descriptions of the three groups of reference sample(s) can be referred to the descriptions related to FIG. 9, and will not be repeated here.

[0189] For example, as shown in FIG. 10, the AI module of device #1 may sample the inputs of the local AI model, the latent layer outputs, and the outputs of the local AI model to obtain three groups of local data sample(s), respectively. The three groups of local data sample(s) correspond to the three groups of reference data sample(s). Then the AI module of the device #1 measures the distances between the local data sample(s) and the reference data sample(s) group by group to obtain three distances corresponding to the three groups, namely distance #1, distance #2 and distance #3 in FIG. 10. If the average distances of these three groups are consistently below a threshold, the AI module of the device #1 may tell that the current inference procedure works as expected, otherwise the AI module may tell it is abnormal.

[0190] FIG.10 is merely an example and shall not constitute any limitation on the present application.

[0191] Further, optionally, the device may also receive information indicating the Q layer(s).

[0192] For example, the information may be Q indicator(s) used to indicate the Q layer(s) related to the Q group(s) of reference data sample(s), respectively.

[0193] As an example, the Q indicator(s) may be the index(s) of the Q group(s) of reference data sample(s).

[0194] Alternatively, the Q layer(s) related to Q group(s) of reference data sample(s) may be predefined.

[0195] Further, optionally, the device may also receive information indicating the condition for determining whether

the inference procedure is normal.

[0196] Alternatively, the condition may be predefined.

[0197] Alternatively, the condition may be determined by the device itself.

[0198] The distance(s) between the Q group(s) of reference data sample(s) and the Q group(s) of local data sample(s) may be measured through the corresponding Q scoring function(s).

[0199] In the case of a plurality of scoring functions, the Q scoring functions may be the same or different.

[0200] Further, optionally, the device may also receive the Q scoring function(s) from the other device.

[0201] Alternatively, the Q scoring function(s) may be predefined.

[0202] Alternatively, the Q scoring function(s) may be determined by the device itself.

[0203] Raw data may be considered as having user privacy. It may be against the privacy policy to transmit raw data. In addition, transmitting raw data may consume a lot of resources. It may be inefficient to transmit raw data.

[0204] The embodiment of the present application provides a communication method where raw data is compressed. Compression is to project high-dimensional data into a low-dimensional one by a transformation.

[0205] The raw data may include the reference data sample(s) mentioned above. For example, the reference data sample(s) may be compressed before being transmitted. Specifically, Q group(s) of the reference data sample(s) may be compressed to a lower dimensional space than the original dimensional space before being transmitted.

[0206] In this way, bandwidth for the reference data sample(s) can be saved and data transmission efficiency can be improved. At the same time, raw data that is the reference data sample(s), can be protected.

[0207] The raw data may include the local data sample(s) mentioned above. The distance(s) between the reference data sample(s) and the local data sample(s) may be replaced by compressed reference data sample(s) and compressed local data sample(s). The technical solution mentioned above can be done with lower dimensional space. For example, the inference cycle detection can be implemented with lower dimensional space. In this way, computational complexity can be reduced which is beneficial to improving processing efficiency. For example, it can be conducive to labeling data in real-time.

[0208] Raw data may be encoded or compressed to a lower dimensional space by a compressor. The encoder can also be called a compressor. The encoder can be linear or non-linear.

[0209] FIG. 11 is a schematic diagram of two examples of encoders.

[0210] For example, the encoder may be a linear encoder realized with some standard basis such as Fourier basis, discrete cosine transform (DCT) or wavelets; Or the encoder maybe a linear encoder realized with some customized basis. For example, these bases may form a unitary matrix or an orthonormal matrix.

[0211] As shown in FIG. 11, the encoder and decoder are aligned on matrix U. Matrix U can be used as a codebook.

For example, matrix U may be a unitary matrix. The encoder may encode the input x through U^H to obtain output c with a lower dimension. c may satisfy the following formula:

[0212] $$c = U^H x .$$

[0213] The decoder can decode c through U to obtain output $\hat{x}$ with the original dimension. $\hat{x}$ may satisfy the following formula:

[0214] $$\hat{x} = Uc .$$

[0215] For another example, the encoder may be a non-linear encoder realized with an AI model, such as DNN. As shown in FIG. 11, the encoder and decoder may be realized with DNNs. The encoder may encode x to c, where c may satisfy the following formula:

[0216] $$c = F(x;\alpha) .$$

[0217] α represents the parameters of the encoder $F(\ )$.

[0218] The decoder may decode c to $\hat{x}$, where $\hat{x}$ may satisfy the following formula:

[0219] $$\hat{x} = G(c;\beta) .$$

[0220] β represents the parameters of the decoder $G(\ )$.

[0221] DNNs can be the approximation of matrix U.

[0222] Unlike the traditional compression schemes built for reliable reconstruction, the encoder in the embodiments of the present application may avoid a reliable reconstruction but preserve as much topological distances as possible, when the data is compressed into a lower dimensional space. That is to say, the relative distance between two data samples in their original dimensional space may be well preserved after being encoded into a low-dimensional space.

[0223] FIG. 12 is a schematic flowchart of a communication method provided by an embodiment of the present application.

[0224] As shown in FIG. 12, a method 1200 includes the following steps.

[0225] Step 1210, a second network element compresses Q group(s) of first raw data sample(s) to obtain Q group(s) of first data sample(s), where Q is a positive integer.

[0226] The Q group(s) of the first data sample(s) is from compressed Q group(s) of first raw data sample(s) which is compressed according to Q transformation matrix(es).

[0227] Step 1220, a first network element receives the Q group(s) of first data sample(s) from the second network element.

[0228] In step 1210, one first data sample is obtained by compressing the corresponding first raw data sample. In other words, the dimension of the first data sample is smaller than the dimension of the corresponding first raw data sample.

[0229] The reference data sample(s) mentioned above is an example of first data sample(s). The compressed reference data sample(s) mentioned above is an example of first raw data sample(s). Method 1200 will be illustrated using this as an example.

[0230] Method 1200 may be applied to an inference cycle of an AI model. Correspondingly, the first raw data sample(s) is related to the inference cycle of AI model(s).

[0231] Optionally, Q group(s) of compressed reference data sample(s) may correspond to Q layer(s) of AI model(s), respectively.

[0232] In other words, Q group(s) of reference data sample(s) may correspond to Q layer(s) of AI model(s), respectively.

[0233] Each group may correspond to one layer of AI model(s). Different groups may correspond to different layers.

[0234] As mentioned above, each group corresponds to output data or input data of one layer of AI model(s).

[0235] The Q layer(s) may belong to one or more AI models.

[0236] The specific description of the corresponding relationship can refer to the previous text, such as FIG. 9 or FIG. 10, and will not be repeated here.

[0237] For example, the second network element may be a network device or a terminal device. The second network element may be the device #2 mentioned above.

[0238] For example, the first network element may be a network device or a terminal device. The first network element may be the device #1 mentioned above.

[0239] According to the above technical solution, the first data sample is a low-dimensional data sample which is compressed according to a transformation matrix. In this way, the bandwidth for the first data sample(s) can be saved and data transmission efficiency can be improved. At the same time, first raw data can be protected.

[0240] The following describe two examples (example#1 and example #2) of compressing the reference data sample.

[0241] Example #1

[0242] Optionally, in step 1210, second network element may compress Q group(s) of reference data sample(s) according to Q first transformation matrix(es) respectively to obtain the Q group(s) of compressed reference data sample(s).

[0243] Each first transformation matrix in the Q first transformation matrix(es) corresponds to one of the Q group(s), respectively. Correspondingly, the Q first transformation matrix(es) may correspond to the Q layer(s), respectively.

[0244] The "first" in "first transformation matrix" is only used to illustrate that the transformation matrix can be used for compressing raw data and does not have any other limiting effect.

[0245] When Q is greater than 1, the Q first transformation matrices corresponding to different groups can be the same or different.

[0246] Optionally, a first transformation matrix be a unitary matrix or an orthonormal matrix. The first transformation matrix can be called basis or reference basis.

[0247] In some embodiments, each basis vector of the first transformation matrix may be a standard basis such as Fourier basis, DCT basis, wavelet basis, or the like.

[0248] In some embodiments, basis vectors of the first transformation matrix may be built as needed. As an example, basis vectors of the first transformation matrix may be built on the distribution of the corresponding group of the reference data samples.

[0249] A raw data sample represented by the first transformation matrix could be written as a finite weighted linear combination of elements of the first transformation matrix. The coefficients of this weighted linear combination are referred to as coordinates of the vector with respect to the first transformation matrix. For example, a compressed reference data sample can be represented by the coefficients with respect to the first transformation matrix.

[0250] In order to facilitate understanding of the embodiment of the present application, the following describes an example process of compression.

[0251] FIG. 13 is a schematic diagram of an example compression process of a reference data sample.

[0252] As shown in FIG. 13, one reference data sample x may be denoted as an $n \times 1$ reference sample, where n is an integer greater than 1. x is taken from the original high-dimensional space. The first transformation matrix U corresponding to the reference data sample x may be denoted as an $n \times r$ matrix, where r is a positive integer smaller than n. U may be a unitary or orthonormal matrix. For the convenience of description, the column is used as a basis vector in the embodiments of the present application. One column of U is one of the basis vectors, which means that any two columns of U are perfectly orthogonal to each other. As shown in FIG. 13, the matrix U consists of r basis vectors. It can be easily applied to that basis matrix whose rows are basis vectors; simply U^H.

[0253] x can be represented by a weighted linear combination of each column of U: $x = Uc$, where c is $r \times 1$ spectrum coefficients or weights. c is an equivalent low-dimensional space data (vector) of x, or in other words, c is the compressed reference data sample of x. Further, r<<n. Matrix U may be a unitary matrix, in which case $U^H U = I$ and $c = U^H x$. The matrix U^H is the encoder or compressor that encodes a high-dimensional ($n \times 1$) reference data sample x into a low-dimensional ($r \times 1$) compressed reference data sample c. In other implementations, U^H can also be considered

as the first transformation matrix. In order to facilitate understanding of the embodiment of the present application, U is taken as the first transformation matrix as an example.

[0254] In order to facilitate understanding of the embodiment of the present application, the following takes Q=2 as an example for explanation. Group #1 of reference data sample(s) may be denoted as $X_1 = \left[x_{1,1} \quad x_{1,2} \quad \cdots \quad x_{1,M_1} \right]$, which may be encoded to a compressed version with the conjugate transpose of the first transformation matrix U_1. $x_{1,1}$ is the first reference data sample in group #1 of reference data sample(s), $x_{1,2}$ is the second reference data sample in group #1 of reference data sample(s), and so on. M_1 is the number of elements in group # 1 of reference data sample(s). The number of reference data samples is the number of compressed reference samples. M_1 is a positive integer. The compressed version is the group #1 of compressed reference data sample(s), which can be denoted as $\mathbb{C}_1 = \left[c_{1,1} \quad c_{1,2} \quad \cdots \quad c_{1,M_1} \right]$. $X_1 = U_1 \mathbb{C}_1$. $c_{1,1}$ is the first compressed reference data sample in group #1 of compressed reference data sample(s), $c_{1,2}$ is the second reference data sample in group #1 of compressed reference data sample(s), and so on. The group #2 of reference data sample(s) may be denoted as $X_2 = \left[x_{2,1} \quad x_{2,2} \quad \cdots \quad x_{2,M_2} \right]$, which may be encoded to a compressed version with the conjugate transpose of the first transformation matrix U_2. $x_{2,1}$ is the first reference data sample in group #2 of reference data sample(s), $x_{2,2}$ is the second reference data sample in group #2 of reference data sample(s), and so on. M_2 is the number of elements in group # 2 of reference data sample(s). M_2 is a positive integer. The compressed version is the group #2 of compressed reference data sample(s), which can be denoted as $\mathbb{C}_2 = \left[c_{2,1} \quad c_{2,2} \quad \cdots \quad c_{2,M_2} \right]$. $X_2 = U_2 \mathbb{C}_2$. $c_{2,1}$ is the first compressed reference data sample in group #2 of compressed reference data sample(s), $c_{2,2}$ is the second reference data sample in group #2 of compressed reference data sample(s), and so on. U_1 and U_2 may be the same or different. In step 1220, the first network element receives $\mathbb{C}_1$ and $\mathbb{C}_2$. Further, the first network element may also receive U_1 and U_2.

[0255] For example, each column of matrix U above may be a standard basis such as Fourier basis, DCT basis, wavelet basis, or the like.

[0256] For another example, the r columns of the matrix U above may be built on the distribution of the corresponding group of the reference data samples.

[0257] An example procedure to calculate the matrix U on the distribution of the corresponding group of the reference data samples may be as follows:

[0258] 1) Accumulating a sufficient amount (M) $n \times 1$ reference data samples: $x_1, x_2, \ldots, x_M$. The M reference data samples belong to the same group. M <<n. M is a positive integer.

[0259] FIG. 14 is a schematic diagram of an example X.

[0260] 2) Juxtaposing the M reference data samples into a $n \times M$ matrix $X = \begin{bmatrix} x_1 & x_2 & \ldots & x_M \end{bmatrix}$. The order of the reference data samples in the matrix X does not matter.

[0261] 3) Applying a rank-reduced singular value decomposition (SVD) on X : $X = U \sum V^H$, where U is $n \times r$ unitary or orthonormal matrix representing a commonality among all the M reference data samples, V^H is a unitary or orthogonal matrix. $\sum$ is a diagonal matrix.

[0262] In some embodiments, the Q first transformation matrix(es) may be determined by the second network element.

[0263] When the second network element is a network device, the Q first transformation matrix(es) may be configured by the network device.

[0264] Optionally, method 1200 may also include: sending information #1 (an example of the first information) indicating the Q first transformation matrix(es) by the second network element to the first network element.

[0265] For example, the information #1 may include one or more first transformation matrices and the correspondence between the one or more first transformation matrices and the Q group(s) of the compressed reference data sample(s).

[0266] For another example, the information #1 may include one or more matrices related to the Q first transformation matrix(es) and the correspondence between the one or more matrices and the Q group(s) of the compressed reference data sample(s), so that the first network element can determine the Q first transformation matrix(es).

[0267] Exemplarily, the second network element may send Q conjugate transpose matrix(es) of the Q first transformation matrix(es).

[0268] For another example, the information #1 may include the index(es) of the Q first transformation matrix(es).

[0269] Exemplarily, there may be multiple first candidate transformation matrices in the first network element. As an example, there may be multiple candidate first transformation matrices with different sizes of space to achieve different resolutions. The multiple candidate first transformation matrices with different sizes of space may be multiple matrices with different numbers of columns. The information #1 may include the index of the Q first transformation matrix(es) within the multiple candidates.

[0270] The information # 1 can also be in other forms, as long as it can indicate which group corresponds to which first transformation matrix.

[0271] In some embodiments, the Q first transformation matrix(es) may be determined by the first network element. The first network element may send information #2 indicating the Q first transformation matrix(es) to the second network element.

[0272] The form of information #2 may refer to the information # 1, and will not be repeated here.

[0273] In some embodiments, the correspondence between the Q first transformation matrix(es) and the Q group(s) may be predefined.

[0274] The following describe the Q layer(s).

[0275] In some embodiments, the Q layer(s) may be determined by the second network element.

[0276] Optionally, method 1200 may also include: sending information #3 (an example of the third information) indicating the correspondence between the Q layer(s) and the Q group(s) by the second network element to the first network element.

[0277] For example, the information #3 may include the Q indicator(s) indicating the Q layer(s) respectively.

[0278] The information # 3 can also be in other forms, as long as it can indicate which group corresponds to which layer.

[0279] In some embodiments, the Q layer(s) may be determined by the first network element. The first network element may send information #4 indicating the Q layer(s) to the second network element.

[0280] The form of information #4 may refer to the information #3, and will not be repeated here.

[0281] In some embodiments, the correspondence between Q layer(s) and Q group(s) may be predefined.

[0282] If the dimensions of the reference data sample are high, the first transformation matrix may also request high dimensions. In addition, if the first transformation matrix is an orthonormal matrix, it cannot be compressed. The first transformation matrix may require high bandwidth, and affect transmission efficiency.

[0283] For example, first transformation matrix U may be denoted as an $n \times r$ matrix. If n is a large number, sending the first transformation matrix may require a lot of resources, which can affect transmission efficiency.

[0284] Example #2

[0285] Optionally, step 1210 may include: sampling Q group(s) of reference data sample(s), by the second network element, through Q sampling matrix(es) respectively to obtain the sampling result(s) of the Q group(s) of reference data sample(s); and compressing sampling result(s) of the Q group(s) of reference data sample(s), by the second network element, according to the Q second transformation matrix(es) respectively to obtain the Q group(s) of compressed reference data

sample(s).

[0286] The sampling matrix may be used to sample values at some positions of an original data example.

[0287] For one reference data sample, the second network element may sample values at some positions of the reference data example through the sampling matrix. Then the second network element compresses the sampling result of the reference data sample according to the second transformation matrix.

[0288] Each sampling matrix in the Q sampling matrix(es) corresponds to one of the Q group(s), respectively. Correspondingly, the Q sampling matrix(es) may correspond to the Q layer(s), respectively.

[0289] Each second transformation matrix in the Q second transformation matrix(es) corresponds to one of the Q group(s), respectively. Correspondingly, the Q second transformation matrix(es) may correspond to the Q layer(s), respectively.

[0290] The "second" in "second transformation matrix" is only used to illustrate that the transformation matrix is related to the compression of the sampling result of the raw data and does not have any other limiting effect. The second transformation matrix can also be called a compact matrix.

[0291] When Q is greater than 1, the Q sampling matrices corresponding to different groups can be the same or different.

[0292] When Q is greater than 1, the Q second transformation matrices corresponding to different groups can be the same or different.

[0293] The following describes the relationship between the first transformation matrix, the sampling matrix and the second transformation matrix.

[0294] Optionally, the Q second transformation matrix(es) may be obtained by sampling the Q first transformation matrix(es) with the Q sampling matrix(es), respectively.

[0295] A first transformation matrix may be sampled to a compact matrix which is smaller than the first transformation matrix through a sampling matrix.

[0296] Optionally, a sampling matrix may be a random matrix or a pseudo-random matrix.

[0297] A first transformation matrix may be $n \times r$ matrix, and the corresponding sampling matrix may be denoted as $m \times n$ matrix. m is a positive integer smaller than n. Further, m<<n. For example, the sampling matrix P may be as follows:

[0298]
$$P = \begin{bmatrix} 0 & \dots & 1 & \dots & 0 & \dots & 0 & \dots & 0 \\ 0 & \dots & 0 & \dots & 1 & \dots & 0 & \dots & 0 \\ & & & & \dots & & & & \\ 0 & \dots & 0 & \dots & 0 & \dots & 1 & \dots & 0 \end{bmatrix}.$$

[0299] Only one position in each row of the sampling matrix has a value other than 0. For example, each row of the sampling matrix has only one "1", and the remaining value(s) in each row are "0". In this way, the position of the value other

than 0 in each row of the sample matrix indicates the sampled position in the raw data sample. Correspondingly, the number of rows in the sampling matrix is the number of positions sampled in the raw data sample.

[0300] The above is merely an example of a sampling matrix. The sampling matrix can also be in other forms.

[0301] In order to facilitate understanding of the embodiment of the present application, the following describes a possible process of the compressing first transformation matrix.

[0302] FIG. 15 is a schematic diagram of an example compression process of a first transformation matrix.

[0303] One reference data sample x may be denoted as an $n \times 1$ sample. A first transformation matrix U corresponding to x may be denoted as an $n \times r$ matrix. A sampling matrix P corresponding to x may be applied to U. P may be denoted as an $m \times n$ matrix, where m<n, and m is a positive integer. Further, m<<n. Each row of P has only one "1" to indicate the position of x to be sampled, and the remaining value(s) in each row are "0". P may be used to "compress" U into a compact matrix θ, which is an $m \times r$ matrix. As shown in FIG. 15, $\theta = PU$ and $x' = \theta c$. x' is an $m \times 1$ sample composed of the values sampled from x. According to the technical solution mentioned above, since m<n, θ is smaller than U. Therefore, θ can be a better alternative to U.

[0304] The following takes two groups mentioned above as an example for explanation. Group #1 of reference data sample(s) may be denoted as $X_1 = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,M_1} \end{bmatrix}$. Group #2 of reference data sample(s) may be denoted as $X_2 = \begin{bmatrix} x_{2,1} & x_{2,2} & \cdots & x_{2,M_2} \end{bmatrix}$. The first transformation matrix U_1 and the first transformation matrix U_2 may be different. The sampling matrix P_1 corresponding to group #1 and the sampling matrix P_2 corresponding to group #2 may be different. U_1 is $n_1 \times r_1$. U_2 is $n_2 \times r_2$. n_1 and n_2 refer to n mentioned above. r_1 and r_2 refer to r mentioned above. If n_1 and/or n_2 are very big numbers, P_1 can be applied to the U_1, and/or P_2 can be applied to U_2. P_1 is $m_1 \times n_1$, each row of which has only one "1" to indicate the position of $x_{1,i}$ to be sampled, and P_2 is $m_2 \times n_2$, each row of which has only one "1" to indicate the position of $x_{2,i}$ to be sampled. P_1 can "compress" U_1 into a second transformation matrix θ_1 of $m_1 \times r_1$ as $\theta_1 = P_1 U_1$. In case of $m_1 \ll n_1$, θ_1 is much smaller than U_1, and θ_1 can be a better alternative to U_1. P_2 can "compress" U_2 into a second transformation matrix θ_2 of $m_2 \times r_2$ as $\theta_2 = P_2 U_2$. In case of $m_2 \ll n_2$, θ_2 is much smaller than U_2, and θ_2 can be a better alternative to U_2.

[0305] When the second network element compresses the Q group(s) of reference data sample(s) with the Q sampling matrix(es) and the Q second transformation matrix(es), the relevant compression method may refer to Example # 4, where the

local data sample may be replaced with reference data sample, and will not be repeated here.

[0306] The second network element may obtain the Q sampling matrix(es) and the Q second transformation matrix(es) in various ways.

[0307] In some embodiments, the Q sampling matrix(es) and the Q second transformation matrix(es) may be predefined.

[0308] In some embodiments, the Q sampling matrix(es) and the Q second transformation matrix(es) may be determined by the second network element.

[0309] For example, the second network element may calculate the Q second transformation matrix(es) through the Q sampling matrix(es) and the Q first transformation matrix(es). The Q first transformation matrix(es) and the Q sampling matrix(es) may be determined by the second network element. As an example, the Q first transformation matrix(es) and the Q sampling matrix(es) may be generated by the second network element.

[0310] In some embodiments, at least one of the Q sampling matrix(es), the Q second transformation matrix(es) or the Q first transformation matrix(es) may be configured by the other network element such as the first network element, while other items that are not configured by the other network element may be predefined or determined by the second network element itself.

[0311] Example #2-1: the second network element may receive the Q sampling matrix(es) and the Q second transformation matrix(es) from other network element.

[0312] Example #2-2: the second network element may receive the Q sampling matrix(es) and Q matrix(es) related to the Q second transformation matrix(es) from other network element, where the Q matrix(es) can be used to calculate the Q second transformation matrix(es). For example, the Q matrix(es) may be Q left inverse matrix(es) of the Q second transformation matrix(es).

[0313] Example #2-3: the second network element may receive the Q sampling matrix(es) and the Q first transformation matrix(es) from other network element. The Q second transformation matrix(es) can be calculated based on the Q sampling matrix(es) and the Q first transformation matrix(es).

[0314] Example #2-4: the second network element may receive the Q first transformation matrix(es) from the other network element. The Q sampling matrix(es) may be generated by the second network element. The Q second transformation matrix(es) can be calculated based on the Q sampling matrix(es) and the Q first transformation matrix(es).

[0315] Example #2-5: the second network element may receive the Q first transformation matrix(es) from the other network element. The Q sampling matrix(es) may be predefined. The Q second transformation matrix(es) can be calculated based on the Q sampling matrix(es) and the Q first transformation matrix(es).

[0316] In addition, the second network element can also determine the Q second transformation matrix(es) through

other methods.

[0317] In example #2, the data sample can be obtained by compressing the raw data sample according to the sampling matrix and the transformation matrix. The dimensions of the sampling matrix and transformation matrix are smaller, which is beneficial to reducing the resources required for transmitting the sampling matrix and transformation matrix, thereby improving transmission efficiency.

[0318] Further, optionally, the method 1200 may also include step 1230.

[0319] Step 1230, the first network element measures the distance(s) between q group(s) of the first data sample(s) in the Q group(s) of the first data sample(s) and q group(s) of the second data sample(s), respectively. q is a positive integer less than or equal to Q.

[0320] The distance between the two in the embodiment of the present application can also be understood as the difference between the two. For example, the distance(s) between q group(s) of the first data sample(s) and q group(s) of the second data sample(s) can also be referred to as the difference(s) between q group(s) of the first data sample(s) and q group(s) of the second data sample(s).

[0321] The local data sample(s) generated by the second network element can be transmitted to first network element as reference data sample(s) for the AI model on the first network element.

[0322] For example, the second network element may transmit group #1 of its local data sample(s) and group #2 of its local data sample(s) to the first network element. The group #1 of the second network element's local data sample(s) can be regarded as the group #1 of the reference data sample(s) transmitted to the device #1. The group #2 of the second network element's local data sample(s) can be regarded as the group #2 of the reference data sample(s) transmitted to the device #1.

[0323] Exemplarily, step 1230 may be executed by the AI module of the first network element.

[0324] The q group(s) of the second data sample(s) corresponds to the q group(s) of the first data sample(s), respectively. The compression method of the q group(s) of the second data sample(s) is related to the compression method of the q group(s) of the first data sample(s).

[0325] In step 1230, one second data sample is obtained by compressing the corresponding second raw data sample. In other words, the dimension of the second data sample is smaller than the dimension of the corresponding second raw data sample.

[0326] The local data sample(s) mentioned above may be an example of second data sample(s). The compressed local data sample(s) mentioned above may be an example of second raw data sample(s). Method 1200 will be illustrated using this as an example.

[0327] Method 1200 may be applied to the inference cycle of AI model(s). Correspondingly, the second data sample(s)

is related to the inference cycle of AI model(s).

[0328] Optionally, q group(s) of compressed local data sample(s) may correspond to q layer(s) of AI model(s), respectively.

[0329] In other words, q group(s) of local data sample(s) may correspond to q layer(s) of AI model(s), respectively.

[0330] Each group may correspond to one layer of AI model(s). Different groups may correspond to different layers.

[0331] As mentioned above, each group corresponds to output data or input data of one layer of AI model(s).

[0332] The q layer(s) may belong to one or more AI models. The method 1200 mainly takes q layer(s) belonging to one AI model as an example.

[0333] The specific description of the corresponding relationship can refer to the previous text, such as FIG. 9 or FIG. 10, and will not be repeated here.

[0334] The following describes two examples (example #3 and example #4) of compressing the local data sample.

[0335] Example #3

[0336] Optionally, the first network element may compress q group(s) of local data sample(s) according to q first transformation matrix(es) respectively to obtain the q group(s) of compressed local data sample(s).

[0337] Each transformation matrix in the q first transformation matrix(es) corresponds to one of the q group(s), respectively. Correspondingly, the q first transformation matrix(es) may correspond to the q layer(s), respectively.

[0338] When q is greater than 1, the q first transformation matrices corresponding to different groups can be the same or different.

[0339] For example, the value of q may be determined by the first network element. Alternatively, the value of q may be indicated by the second network element. Alternatively, the value of q may be predefined.

[0340] The following takes q=2 as an example for explanation. The group #1 of local data sample(s) may be denoted as $\hat{X}_1 = \left[\hat{x}_{1,1} \quad \hat{x}_{1,2} \quad \cdots \quad \hat{x}_{1,K_1} \right]$. $\hat{x}_{1,1}$ is the first local data sample in the group #1 of local data sample(s), $\hat{x}_{1,2}$ is the second local data sample in the group #1 of local data sample(s). The K_1 local data sample(s) may be obtained by randomly sampling K_1 data sample(s) on the corresponding layer #1. For example, the corresponding layer #1 may be the layer indicated by the indicator with the group #1 of compressed reference data sample(s). K_1 is positive integer. The K_1 data sample(s) may be the input(s) or output(s) of the corresponding layer #1. This is merely an example. The embodiments of the present application do not limit this. For example, the first network element may sample each data sample on the corresponding layer #1. Then the compressed local data sample $\hat{c}_{1,i}$ may be calculated as $\hat{c}_{1,i} = U_1^{+} \hat{x}_{1,i}$. U_1^{+} is the reverse of U_1.

The group #1 of compressed local data sample(s) is denoted as $\hat{\mathbb{C}}_1 = \left[\hat{c}_{1,1} \quad \hat{c}_{1,2} \quad \dots \quad \hat{c}_{1,K_1} \right]$.The group #2 of local data sample(s) may be denoted as $\hat{X}_2 = \left[\hat{x}_{2,1} \quad \hat{x}_{2,2} \quad \dots \quad \hat{x}_{2,K_2} \right]$. $\hat{x}_{2,1}$ is the first local data sample in the group #2 of local data sample(s), $\hat{x}_{2,2}$ is the second local data sample in the group #2 of local data sample(s). The K_2 local data sample(s) may be obtained by randomly sampling K_2 data sample(s) on the corresponding layer #2. K_2 is positive integer. For example, the corresponding layer #2 may be the layer indicated by the indicator with the group #2 of compressed reference data sample(s). The K_2 data sample(s) may be the input(s) or output(s) of the corresponding layer #2. This is merely an example. The embodiments of the present application do not limit this. For example, the first network element may sample each data sample which may be the input(s) or output(s) of the corresponding layer. Then the compressed local data sample $\hat{c}_{2,i}$ may be calculated as $\hat{c}_{2,i}=U_2^{+}\hat{x}_{2,i}$. U_2^{+} is the reverse of U_2.The group #2 of compressed local data sample(s) is denoted as $\hat{\mathbb{C}}_2 = \left[\hat{c}_{2,1} \quad \hat{c}_{2,2} \quad \dots \quad \hat{c}_{2,K_2} \right]$.

[0341] The specific compression method may refer to Example # 1, where the reference data sample may be replaced with a local data sample, and will not be repeated here.

[0342] The q first transformation matrix(es) may be related to the q group(s) of compressed reference data sample(s). For example, the q first transformation matrix(es) may also be used to compress the q group(s) of reference data sample(s), respectively.

[0343] The q first transformation matrix(es) belongs to the Q first transformation matrix(es). The determination method of the Q first transformation matrix(es) may refer to Example # 1.

[0344] The q layer(s) belongs to the Q layer(s). The determination method of the Q layer(s) may refer to Example # 1.

[0345] Example #4

[0346] Optionally, the first network element may sample q group(s) of local data sample(s) through q sampling matrix(es) respectively to obtain the sampling result(s) of the q group(s) of local data sample(s); the first network element compresses sampling result(s) of the q group(s) of local data sample(s) according to q second transformation matrix(es) respectively to obtain the q group(s) of compressed local data sample(s).

[0347] For one local data sample, the first network element may sample values at some positions of the local data example through the sampling matrix. Then the first network element compresses the sampling result of the local data sample according to the second transformation matrix.

[0348] Each sampling matrix in the q sampling matrix(es) corresponds to one of the q group(s), respectively.

Correspondingly, the q sampling matrix(es) may correspond to the q layer(s), respectively.

[0349] Each second transformation matrix in the q second transformation matrix(es) corresponds to one of the q groups, respectively. Correspondingly, the q second transformation matrix(es) may correspond to the q layer(s), respectively.

[0350] When q is greater than 1, the q sampling matrices corresponding to different groups can be the same or different.

[0351] When q is greater than 1, the q second transformation matrices corresponding to different groups can be the same or different.

[0352] The following takes q=2 as an example for explanation. The group #1 of local data sample(s) may be denoted as $\hat{X}_1 = \left[\hat{x}_{1,1} \quad \hat{x}_{1,2} \quad \dots \quad \hat{x}_{1,K_1}\right]$. The relevant description of group # 1 local data sample(s) can be referred to Example #3 and will not be repeated here. The first network element samples the group #1 of local data sample(s), where the first network element may sample the m_1 position(s) indicated by the sampling matrix #1 P_1 in the local data sample $\hat{x}_{1,i}$ into a $m_1 \times 1$ local sample $\hat{x}'_{1,i}$. m_1 is a positive integer. $m_1 \leq n_1$. n_1 is the dimension of a local data sample in the group #1. Then the compressed local data sample $\hat{c}_{1,i}$ may be calculated as $\hat{c}_{1,i} = \theta_1^{+} \hat{x}'_{1,i}$. The group #1 of compressed local data sample(s) is denoted as $\hat{\mathbb{C}}_1 = \left[\hat{c}_{1,1} \quad \hat{c}_{1,2} \quad \dots \quad \hat{c}_{1,K_1}\right]$. The group #2 of local data sample(s) may be denoted as $\hat{X}_2 = \left[\hat{x}_{2,1} \quad \hat{x}_{2,2} \quad \dots \quad \hat{x}_{2,K_2}\right]$. The relevant description of group # 2 of local data sample(s) can be referred to example #3 and will not be repeated here. The first network element samples the group #2 of local data sample(s), where the first network element may sample the m_2 position(s) indicated by the sampling matrix #2 P_2 in the local data sample $\hat{x}_{2,i}$ into a $m_2 \times 1$ local sample $\hat{x}'_{2,i}$. m_2 is a positive integer. $m_2 \leq n_2$. n_2 is the dimension of a local data sample in the group #2. Then the compressed local data sample $\hat{c}_{2,i}$ may be calculated as $\hat{c}_{2,i} = \theta_2^{+} \hat{x}'_{2,i}$. The group #2 of compressed local data sample(s) is denoted as $\hat{\mathbb{C}}_2 = \left[\hat{c}_{2,1} \quad \hat{c}_{2,2} \quad \dots \quad \hat{c}_{2,K_2}\right]$.

[0353] The q sampling matrix(es) and the q second transformation matrix(es) may be related to the q group(s) of compressed reference data sample(s). For example, the q sampling matrix(es) and the q second transformation matrix(es) may also be used to compress the q group(s) of reference data sample(s), respectively. For another example, q first transformation matrix(es) may be used to compress the q group(s) of reference data sample(s), respectively, where the q first transformation matrix(es) may also be used to calculate the q second transformation matrix(es).

[0354] As mentioned above, the first network element may multiply the sampling result(s) of the q group(s) of local data sample(s) with the left inverse of the q second transformation matrix(es) to obtain the q group(s) of compressed local data

sample(s).

[0355] The first network element may obtain the left inverse of the q second transformation matrix(es), such as θ_1^+ and θ_2^+ mentioned above in various ways.

[0356] In some embodiments, the Q sampling matrix(es) and the Q second transformation matrix(es) may be predefined. The first network element calculates the left inverse of the Q second transformation matrix(es).

[0357] For example, the θ_1 and θ_2 may be predefined. And the first network element left inverses θ_1 into θ_1^+ and θ_2 into θ_2^+.

[0358] Alternatively, the Q sampling matrix(es) and the left inverse of Q second transformation matrix(es) may be predefined.

[0359] In some embodiments, the Q sampling matrix(es) and the Q second transformation matrix(es) may be determined by the first network element. The first network element calculates the left inverse of the Q second transformation matrix(es).

[0360] For example, the first network element may calculate the Q second transformation matrix(es) through the Q sampling matrix(es) and the Q first transformation matrix(es). The Q first transformation matrix(es) and the Q sampling matrix(es) may be determined by the first network element. For example, the Q first transformation matrix(es) and the Q sampling matrix(es) may be generated by the first network element.

[0361] And the first network element may indicate the Q sampling matrix(es) and the Q second transformation matrix(es) to the second network element. Relevant descriptions may refer to Example # 2.

[0362] In some embodiments, at least one of the Q sampling matrix(es), the Q second transformation matrix(es) or the Q first transformation matrix(es) may be configured by the second network element, while other items that are not configured by the second network element may be predefined or determined by the first network element itself.

[0363] The first network element may receive information#5 (an example of the first information) indicating the left inverse of the Q second transformation matrix(es) from the second network element. The left inverse of the Q second transformation matrix(es) can be calculated through the Q second transformation matrix(es). Thus, the information#5 can also be understood as indicating Q second transformation matrix(es).

[0364] The following describes some example forms of information #5.

[0365] Example #4-1: the information #5 may include the Q sampling matrix(es) and the Q second transformation matrix(es). The first network element calculates the left inverse of the Q second transformation matrix(es).

[0366] For example, the first network element may receive P_1, θ_1, P_2 and θ_2 mentioned above from the second network element, then left inverse the θ_1 into θ_1^+ and θ_2 into θ_2^+.

[0367] Example #4-2: the information #5 may include Q sampling matrix(es) and Q matrix(es) related to the Q second transformation matrix(es), where the Q matrix(es) can be used to determine the left reverse of the Q second transformation matrix(es).

[0368] As an example, the information #5 may include Q sampling matrix(es) and the left reverse of the Q second transformation matrix(es).

[0369] For example, the first network element may receive P_1, θ_1^+, P_2 and θ_2^+ mentioned above from the second network element.

[0370] Example #4-3: the information #5 may include Q sampling matrix(es) and Q first transformation matrix(es). The left inverse of the Q second transformation matrix(es) can be calculated based on the Q sampling matrix(es) and Q first transformation matrix(es).

[0371] For example, the first network element may receive P_1, U_1, P_2 and U_1 mentioned above from the second network element. Then first network element calculates θ_1^+ as $\theta_1^+ = (P_1U_1)^+$ and θ_2^+ as $\theta_2^+ = (P_2U_2)^+$.

[0372] Example #4-4: the information #5 may include Q first transformation matrix(es). The left inverse of the Q second transformation matrix(es) can be calculated based on the Q sampling matrix(es) and Q first transformation matrix(es). The Q sampling matrix(es) may be generated by the first network element. Or the Q sampling matrix(es) may be predefined.

[0373] For example, the first network element may receive U_1 and U_1 mentioned above from the second network element. P_1 and P_2 may be generated locally by the first network element. Then first network element calculates θ_1^+ as $\theta_1^+ = (P_1U_1)^+$ and θ_2^+ as $\theta_2^+ = (P_2U_2)^+$.

[0374] In addition, the first network element can also determine the left reverse of the Q second transformation matrix(es) through other methods. For example, the information #5 may include the index of the matrices mentioned above. Exemplarily, there may be multiple candidate sampling matrices and candidate second transformation matrices in the first network element. The information #5 may include the index of the Q sampling matrix(es) and the index of the Q second transformation matrix(es) within the multiple candidates.

[0375] In addition, the example #3 can also be executed through the Example #4. The first network element doesn't sample value(s) from the local data sample(s), mathematically the sampling matrix being an identity matrix. For example, P_1

is an identity matrix I and P_2 is an identity matrix I. The first network element calculates the left inverse of the second transformation matrix as $\theta_1^+ = (P_1 U_1)^+ = U_1^+$ and $\theta_2^+ = (P_2 U_2)^+ = U_2^+$. If U_1 is unitary, $\theta_1^+ = U_1^+ = U_1^H$. If U_2 is unitary, $\theta_2^+ = U_2^+ = U_2^H$.

[0376] In example #4, the data sample can be obtained by compressing the raw data sample according to the sampling matrix and the second transformation matrix. The dimensions of the sampling matrix and the second transformation matrix are smaller, which is beneficial to reducing the resources required for transmitting the sampling matrix and second transformation matrix, thereby improving transmission efficiency. For example, the second network element may send Q sampling matrix(es) and Q second transformation matrix(es) to the first network element. Compared to sending Q first transformation matrix(es), this way may require fewer transmission resources due to the smaller dimensions of the second transformation matrix and sampling matrix compared to the first transformation matrix, which is beneficial to ensuring transmission efficiency.

[0377] The following describes the distance(s) between the q group(s) of first data sample(s) and the q group(s)of second data sample(s).

[0378] For a compressed local data sample and a compressed reference data sample corresponding to the same layer, the distance between the compressed local data sample and the compressed reference data sample is approximately the same as the distance between the raw local data sample and the raw reference data sample.

[0379] FIG. 16 is a schematic diagram of an example distance on the low spectrum space.

[0380] For example, as shown in FIG. 16, the distance between a local data sample $\hat{x}$ and a reference data sample x may be denoted as $\delta = d(x, \hat{x})$, and the distance between the compressed local data sample $\hat{c}$ and the compressed reference data sample c may be denoted as $\delta = d(c, \hat{c})$, where $d(\ )$ is the scoring function. $d(x, \hat{x}) \approx U d(c, \hat{c})$.

[0381] Therefore, in some scenarios, the distance(s) between the q group(s) of compressed reference data sample(s) and the q group(s)of compressed local data sample(s) can be used to indicate the trend of the distance(s) between the q group(s) of reference data sample(s) and the q group(s) of local data sample(s). The q group(s) of the local data sample(s) may be the input(s) or output(s) of the corresponding layer(s). For example, each group of the local data sample(s) may be obtained by sampling the input(s) or output(s) of the corresponding layer. Further, each group of the local data sample(s) may be obtained by sampling the input(s) or output(s) of the corresponding layer.

[0382] The distance(s) between the q group(s) of the compressed reference data sample(s) and q group(s) of the compressed local data sample(s) may be calculated with q scoring function(s), respectively, where each scoring function of the q scoring function(s) may be used to measure the distance between the compressed local data sample from the group of

compressed local data sample(s) corresponding to the scoring function and a compressed reference data sample from the group of compressed reference data sample(s) corresponding to the scoring function, or each scoring function of the q scoring function(s) may be used to measure the distance between the distribution of the group of compressed local data sample(s) corresponding to the scoring function and the distribution of the group of compressed reference data sample(s) corresponding to the scoring function.

[0383] The q scoring function(s) may correspond to the q group(s), respectively.

[0384] The following describes the q scoring functions.

[0385] The q scoring function(s) may correspond to the q layer(s), respectively.

[0386] When q >1, the q scoring function(s) may be the same or different.

[0387] The first network element may determine the q scoring function(s) in various ways.

[0388] Further, optionally, the method 1200 may also include: the first network element may receive information #6 (an example of the fourth information) indicating the Q scoring function(s) from the second network element. The Q scoring function(s) includes the q scoring function(s). The Q scoring function(s) may correspond to the Q layer(s), respectively.

[0389] For example, the information #6 may include the Q scoring function(s).

[0390] For another example, the information #6 may include the index of the Q scoring function(s).

[0391] Alternatively, the first network element may get the q scoring function(s) through other methods. For example, the q scoring function(s) corresponding to the q layer(s) may be predefined. For another example, the q scoring function(s) corresponding to the q layer(s) may be determined by the first network element.

[0392] In some embodiments, each scoring function may be used to measure the distance between two samples.

[0393] As an example, the scoring function may be one of dot product, inner product, Euclidean distance, and so on.

[0394] As another example, the scoring function may be DNN-based.

[0395] The following takes two groups mentioned above as examples for explanation. The group #1 of compressed reference data sample(s) may be denoted as $\mathbb{C}_1 = \left[c_{1,1} \quad c_{1,2} \quad \dots \quad c_{1,M_1} \right]$. The group #2 of compressed reference data sample(s) may be denoted as $\mathbb{C}_2 = \left[c_{2,1} \quad c_{2,2} \quad \dots \quad c_{2,M_2} \right]$.The group #1 of compressed local data sample(s) may be denoted as $\hat{\mathbb{C}}_1 = \left[\hat{c}_{1,1} \quad \hat{c}_{1,2} \quad \dots \quad \hat{c}_{1,K_1} \right]$, where K_1 is the number of the compressed local data samples in the group #1 of compressed local data sample(s) and K_1 is a positive integer. $\hat{c}_{1,1}$ represents the first element in the group #1 of compressed local data sample(s), and $\hat{c}_{1,2}$ represents the second element in the group #1 of compressed local data sample(s), and so on. The group #2 of compressed local data sample(s) may be denoted as $\hat{\mathbb{C}}_2 = \left[\hat{c}_{2,1} \quad \hat{c}_{2,2} \quad \dots \quad \hat{c}_{2,K_2} \right]$, where

K_2 is the number of compressed local data samples in the group #2 of compressed local data sample(s) and K_2 is a positive integer. $\hat{c}_{2,1}$ represents the first element in the group #2 of compressed local data sample(s), and $\hat{c}_{2,2}$ represents the second element in the group #2 of compressed local data sample(s), and so on. There are two scoring functions, namely the scoring function #1 $d_1(\ )$ corresponding to the group #1 and the scoring function # 2 corresponding to the group #2 $d_2(\ )$. The scoring function #1 $d_1\left(c_{1,i},\hat{c}_{1,i}\right)$ is used to measure the distance between two samples $c_{1,i}$ and $\hat{c}_{1,i}$. The scoring function #2 $d_2\left(c_{2,i},\hat{c}_{2,i}\right)$ is used to measure the distance between two samples $c_{2,i}$ and $\hat{c}_{2,i}$. The scoring function #1 $d_1(\ )$ and the scoring function #2 $d_2(\ )$ may be the same or different.

[0396] The distance between each two corresponding groups may be based on the distance between the data samples in the two groups.

[0397] As an example, the distance between each two corresponding groups may be the average minimum distance between the data samples in the two groups.

[0398] The following takes two groups mentioned above as examples for explanation.

[0399] For example, the scoring function #1 $d_1(\ )$ may be used to measure the distance between two samples for group #1. The distance δ_1 between the group #1 of compressed local data sample(s) and the group #1 of compressed reference data sample(s) may be the average minimum distance for the group #1, that is, $\delta_1 = \frac{\sum_{k=1}^{k=K_1} \min_{j=1,2,\ldots,M_1} \left(d_1\left(\hat{c}_{1,k},c_{1,j}\right)\right)}{K_1}$.

The scoring function #2 $d_2(\ )$ may be used to measure the distance between two samples for group #2. The distance δ_2 between the group #2 of compressed local data sample(s) and the group #2 of compressed reference data sample(s) may be the average minimum distance for the group #2, that is $\delta_2 = \frac{\sum_{k=1}^{k=K_2} \min_{j=1,2,\ldots,M_2} \left(d_2\left(\hat{c}_{2,k},c_{2,j}\right)\right)}{K_2}$.

[0400] In some embodiments, each scoring function may be used to measure the distance between two distributions.

[0401] As an example, the scoring function may be one of the following: mutual information, Hilbert-Schmidt independence criterion (HSIC) metric, Kullback-Leibler divergence (KL divergence), graph edit distance, Wasserstein distance, Jensen-Shanon distance (JSD distance), and so on.

[0402] As another example, the scoring function may be DNN-based.

[0403] The following takes two groups mentioned above as examples for explanation.

[0404] There are two scoring functions, namely the scoring function #1 $d_1(\ )$ corresponding to the group #1 and the scoring function # 2 corresponding to the group #2 $d_2(\ )$. The scoring function #1 $d_1\left(\mathbb{C}_1, \hat{\mathbb{C}}_1\right)$ is used to measure the distance between two distributions $\mathbb{C}_1$ and $\hat{\mathbb{C}}_1$ of the group #1. The scoring function #2 $d_2\left(\mathbb{C}_2, \hat{\mathbb{C}}_2\right)$ is used to measure the distance between two distributions $\mathbb{C}_2$ and $\hat{\mathbb{C}}_2$ of the group #2. The scoring function #1 $d_1(\ )$ and the scoring function #2 $d_2(\ )$ may be the same or different.

[0405] The distance between each two corresponding groups may be based on the distance between two distributions of the two groups.

[0406] The following takes two groups mentioned above as examples for explanation.

[0407] For example, the scoring function #1 $d_1(\ )$ may be used to measure the distance between two distributions for the group #1. The distance δ_1 between the group #1 of compressed local data sample(s) and the group #1 of compressed reference data sample(s) may be the distance between two distributions for the group #1, that is, $\delta_1 = d_1\left(\mathbb{C}_1, \hat{\mathbb{C}}_1\right)$. The scoring function #2 $d_2(\ )$ may be used to measure the distance between two distributions for the group #2. The distance δ_2 between the group #2 of compressed local data sample(s) and the group #2 of compressed reference data sample(s) may be the distance between two distributions for the group #2, that is, $\delta_2 = d_2\left(\mathbb{C}_2, \hat{\mathbb{C}}_2\right)$.

[0408] The measure methods of distance for different groups can be the same or different. For example, the distance δ_1 between the group #1 of compressed local data sample(s) and the group #1 of compressed reference data sample(s) may be the average minimum distance for the group #1, and the distance δ_2 between the group #2 of compressed local data sample(s) and the group #2 of compressed reference data sample(s) may be the distance between two distributions for the group #2.

[0409] Optionally, the first network element may calculate the higher order such as root mean square (RMS), standard deviation of δ_1 and δ_2.The higher order is conducive to more accurate determination of the difference between the group of the compressed local data samples and the group of the compressed reference samples.

[0410] For a first data sample and a second data sample corresponding to the same layer, the distance between the first data sample and the second data sample is approximately the same as the distance between the first raw data sample and the

second raw data sample. In this way, computational complexity can be reduced, which is beneficial to improving processing efficiency.

[0411] The first network element may process and/or communicate based on the distance(s) between q group(s) of the first data sample(s) in the Q group(s) of the first data sample(s) and q group(s) of the second data sample(s).

[0412] Optionally, the first network element may send information #7 (an example of the second information) indicating the distance(s) between q group(s) of the first data sample(s) in the Q group(s) of the first data sample(s) and q group(s) of the second data sample(s).

[0413] Exemplarily, information #7 may be transmitted by the communication module of the first network element.

[0414] As an example, the information #7 may indicate the q distance(s) corresponding to the q group(s). For example, the information #7 may include the q distance(s).

[0415] As mentioned before, q is less than or equal to Q. When q is less than Q, the number of groups of compressed reference data samples received by the first network element is greater than the number of distances sent by the first network element.

[0416] The first network element may send the distance(s) in broadcast, multicast, or unicast way.

[0417] If the first network element sends distances of multiple groups, the sending way for distances of different groups can be the same or different.

[0418] As another example, there may be multiple distance ranges. Each distance range corresponds to a level. The information #7 may indicate q level(s) corresponding to the distance range(s) to which the q distance(s) belong.

[0419] As another example, the information #7 may indicate the statistical value of the q distances.

[0420] Exemplarily, the statistical value of the q distances may include the average, maximum, total, or minimum value of the q distances.

[0421] For example, the first network element may send the maximum distance of the q distances.

[0422] The following describes an example explanation of the timing of sending the information #7.

[0423] For example, the first network element may send the information #7 once the distance(s) have been measured.

[0424] For another example, the first network element may send the information #7 in response to the request sent by the other network element(s) for the measurement result.

[0425] For another example, the first network element may send the information #7 when the new measurement result is different from the older measurement result.

[0426] Group #1 is taken as an example. The first network element receives group #1 of compressed reference data sample(s) at time # 1 and calculates the distance based on the current group # 1 of compressed local data sample(s). The first

network element receives group #1 of compressed reference data sample(s) at time # 2 and calculates the distance based on the current group # 1 of compressed local data sample(s). Time # 2 and time # 1 may belong to the same inference cycle of an AI model, and time # 2 is later than time # 1. The first network element may be moving, local data samples may change. Correspondingly, the distances corresponding to group # 1 calculated at different times may also be different. The first network element may send the information #7 when the new measurement result corresponding to time #2 is different from the older measurement result corresponding to time #1.

[0427] In addition, the communication system of the device may receive the new groups of compressed reference data samples, new encoders, and/or new scoring functions from one period of time to another. The AI module of the device may use the most recent compressed reference data samples, encoders, and/or scoring functions to its local data samples and the communication system of the device may transmit the information indicating the most recent measurement results with the most recent compressed reference data samples, encoders, and/or scoring functions to its local data samples.

[0428] Optionally, the first network element may use the distance(s) between q group(s) of the first data sample(s) in the Q group(s) of the first data sample(s) and q group(s) of the second data sample(s) as judgment benchmark in some application scenarios.

[0429] The distance(s) can be used for performing checking.

[0430] Performing checking may include checking whether the current inference cycle is abnormal or not.

[0431] In the embodiment of the application, "checking whether the current inference cycle is abnormal or not " can also be replaced by the following description: checking whether the AI model can work as expected; checking whether the distance(s) meets the expectation; checking whether the distance(s) meets the conditions; checking whether the distance(s) is within the predefined range; checking whether the AI model meets expectation; checking whether the AI model is a candidate model matching another AI model, and so on.

[0432] For the convenience of description, the embodiment of the present application mainly takes checking whether the current inference cycle is abnormal or not.

[0433] In some application scenarios, the measure results may be used to detect whether the current inference cycle is abnormal or not. The detection method can refer to the previous text, replacing the distance(s) in the original dimensional space with the distance(s) in a lower dimensional space, and will not be repeated here.

[0434] Further, optionally, the detection results of the inference cycle may be indicated to another network element.

[0435] In addition, the above actions executed by the second network element can also be executed by the first network element. The above actions executed by the second network element can also be executed by a third network element. The third network element and second network element can be the same device or different devices. The first network element and second

network element are different devices.

[0436] The following is an example of Q=2, which does not constitute a limitation on the technical solution of the present application. Other descriptions can refer to the previous text and will not be repeated here.

[0437] For example, the communication module of the first network element transmits U_1 , group#1 of its compressed local data samples, U_2 and group #2 of its compressed local data samples to the third network element. The third network element receives group#1 of first network element's compressed local data samples as its group #1 of the reference data samples and group #2 of first network element's compressed local data samples as its group #2 of the reference data samples.

[0438] Alternatively, the communication module of the first network element transmits θ_1 , P_1 , group #1 of its compressed local data samples, θ_2 , P_2 , and group #2 of its compressed local data samples to the third network element. The third network element receives group#1 of first network element's compressed local data samples as its group #1 of the reference data samples and group #2 of first network element's compressed local data samples as its group #2 of the reference data samples.

[0439] Alternatively, the communication module of the first network element transmits θ_1^+ , P_1 , and group #1 of its compressed local data samples to the third network element. the communication module of the first network element transmits θ_2^+ , P_2 , and group #2 of its compressed local data samples to the third network element. The third network element receives group#1 of first network element's compressed local data samples as its group #1 of the reference data samples and group #2 of first network element's compressed local data samples as its group #2 of the reference data samples.

[0440] The communication module of the first network element may transmit the scoring function #1 $d_1\left(c_{1,i}, \hat{c}_{1,i}\right)$ that measures the distance between two samples, $c_{1,i}$ and $\hat{c}_{1,i}$, of the group #1. The communication module of the first network element may transmit the scoring function #2 $d_2\left(c_{2,i}, \hat{c}_{2,i}\right)$ that measures the distance between two samples, $c_{2,i}$ and $\hat{c}_{2,i}$, of the group #2. The scoring function #1 $d_1(\ )$ and the scoring function #2 $d_2(\ )$ may be the same or different.

[0441] The scoring function #1 $d_1(\ )$ and the scoring function #2 $d_2(\ )$ may be dot product, inner product, Euclidean distance, and so on. Alternatively, the scoring function #1 $d_1(\ )$ and the scoring function #2 $d_2(\ )$ may be DNN-based.

[0442] Alternatively, the communication module of the first network element may transmit the scoring function #1

$d_1\left(\mathbb{C}_1, \hat{\mathbb{C}}_1\right)$ that measures the distance between two distributions, $\mathbb{C}_1$ and $\hat{\mathbb{C}}_1$, of the group #1. The communication module of the first network element may transmit the scoring function #2 $d_2\left(\mathbb{C}_2, \hat{\mathbb{C}}_2\right)$ that measures the distance between two distributions, $\mathbb{C}_2$ and $\hat{\mathbb{C}}_2$, of the group #2. The scoring function #1 $d_1(\ )$ and the scoring function #2 $d_2(\ )$ may be the same or different.

[0443] The scoring function #1 $d_1(\ )$ and the scoring function #2 $d_2(\ )$ may be mutual information, HSIC metric, KL divergence, graph edit distance, Wasserstein distance, JSD distance, and so on. Alternatively, the scoring function #1 $d_1(\ )$ and the scoring function #2 $d_2(\ )$ may be DNN-based.

[0444] The following describes an exemplary explanation of method 1200 of the embodiments in the present application based on two examples (Example scenario-1 and Example scenario-2).

[0445] Example scenario-1

[0446] Optionally, method 1200 may be used to check AI model generalization. In other words, the method 1200 can be used to check whether the AI model can work.

[0447] For example, the AI module of the first network element may check if the distance(s) satisfies the conditions above. If the AI module of the first network element suspects the distance(s) do not meet the conditions above, it may decide that the AI model cannot work.

[0448] Further, optionally, the method 1200 may also include the following step.

[0449] The first network element may send information #7 indicating the distance(s) between q group(s) of the first data sample(s) and q group(s) of the second data sample(s).

[0450] If the first network element reports the distance(s) between q group(s) of the first data sample(s) and q group(s) of the second data sample(s) to the second network element, the second network element may determine whether the AI model can work.

[0451] Further, if the inference cycle of current AI model deployed on the first network element is abnormal, the current AI model may be replaced. For example, the current AI model may be switched to other AI models. Alternatively, the current AI model may be replaced by a non-AI model.

[0452] The switched model can be configured by the second network element.

[0453] Alternatively, the switched model can also be determined by the first network element and notified to the second network element.

[0454] Example scenario-2

[0455] In some scenarios, a plurality of AI models deployed on different devices may need to work together. These AI models may be trained independently by different providers.

[0456] Optionally, method 1200 may be used to check the interconnection of a plurality of AI models.

[0457] For example, the AI module of the first network element may check if the distance(s) satisfies the conditions above. If the AI module of the first network element suspects the distance(s) do not meet the conditions above, it may decide that the AI model cannot work with another AI model.

[0458] The first network element may send information #7 indicating the distance(s) between q group(s) of the first data sample(s) and q group(s) of the second data sample(s).

[0459] If the first network element reports the distance(s) between q group(s) of the first data sample(s) and q group(s) of the second data sample(s) to the second network element, it can also be performed by the second network element to determine whether the AI model can work with another AI model.

[0460] For example, an encoder and a decoder deployed on different devices may need to work together. The encoder can be deployed on the transmitter side and the decoder can be deployed on the receiver side. The transmitter side is an encoding device. The receiver side is a decoding device. The encoder of the encoding device may output to the decoder of the decoding device.

[0461] The method 1200 may be applied to check whether the encoder and the decoder deployed on different devices can work together.

[0462] The following takes a DNN-based autoencoder as an example. The encoder can be an encoding DNN and the decoder can be a decoding DNN.

[0463] There are two devices, i.e. device #1 and device #2. The AE#1 deployed on the device #1 and AE#2 deployed on the device #2 need to work together. For example, the device #1 may include the modules shown in FIG. 3, where the sensing module may be used to collect the local data, AI module may be used to perform inference on an its local data with encoding DNN #1 in the AE #1, and communication module may be used to receive signals and/or data and transmit signals and/or data. The device #2 may include the modules shown in FIG. 3, where the sensing module may be used to collect the local data, AI module may be used to perform inference on the data received from the encoding DNN on other device with decoding DNN #2 in the AE #2, and communication module may be used to receive signals and/or data and transmit signals and/or data.

[0464] The encoding DNN on the device #1 need to work with the decoding DNN on the device #2. The distance(s) can be used to determine whether the AI models on two devices can work together.

[0465] Exemplarily, the device #1 can be the first network element, and the device #2 can be the second network

element. Alternatively, the device #1 can be the second network element, and the device #2 can be the first network element.

[0466] FIG. 17 is a schematic diagram of the autoencoder with one group of reference data samples.

[0467] For example, as shown in FIG. 17, the device #1 can be the first network element, and the device #2 can be the second network element.

[0468] The relationship between the input to the AE #2 X_{in} and the latent layer output X_{latent} can be represented as $X_{latent} = f_1(X_{in}; \gamma_1)$. $f_1(\ )$ represents the encoder#2 of the AE #2, and γ_1 represents parameters of the encoder#2 $f_1(\ )$. The relationship between the output of the AE #2 X_{out} and the latent layer output can be represented as $X_{out} = g_1(X_{latent}; \varphi_1)$. $g_1(\ )$ represents the decoder #2 of the AE #2, and φ_1 represents parameters of the decoder #2 $g_1(\ )$. X_{latent} is the output of the encoder #2, and also the input of the decoder #2.

[0469] The relationship between the input to the AE #1 $\hat{X}_{in}$ and the latent layer output $\hat{X}_{latent}$ of the AE #1 can be represented as $\hat{X}_{latent} = f_2(\hat{X}_{in}; \gamma_2)$. $f_2(\ )$ represents the encoder #1 of the AE #1, and γ_2 represents parameters of the encoder #1 $f_2(\ )$. The relationship between the output of the AE #1 $\hat{X}_{out}$ and the latent layer output $\hat{X}_{latent}$ of the AE #1 can be represented as $\hat{X}_{out} = g_2(\hat{X}_{latent}; \varphi_2)$. $g_2(\ )$ represents the decoder #1 of the AE #1, and φ_2 represents parameters of the decoder #1 $g_2(\ )$. $\hat{X}_{latent}$ is the output of the encoder #1, and also the input of the decoder #1.

[0470] Method 1200 can be used to check whether AE #1 and AE #2 can work together. For example, method 1200 can be used to check whether the encoder #1 can work with decoder #2.

[0471] The AI module of the device #2 may compress the reference data samples to obtain the compressed reference data samples. The reference data samples may be sampled from X_{latent}.

[0472] The communication module of the device #2 send the compressed reference data samples $\mathbb{C}$ to the device #1.

[0473] Further, the AI module of the device #2 may also generate second transformation matrix θ and sampling matrix P. The communication module of the device #2 may transmit $(\mathbb{C}, \theta, P, d(\ ))$ to the device #1. $d(\ )$ is the scoring function used to measure the distance(s) between the compressed local data samples and the compressed reference data samples.

[0474] The AI module of the device #1 may compress the local data samples to obtain the compressed local data samples according to the second transformation matrix θ and sampling matrix P. The local data samples may be sampled from

$\hat{X}_{latent}$.

[0475] The AI module of the device #1 measures the distance(s) between the reference data samples $\mathbb{C}$ received by the communication module of the device #1 and the local data samples $\hat{\mathbb{C}}$ through the scoring function $d(\ )$.

[0476] The device #1 may check whether the encoder #1 can work with decoder #2 according to the distance(s).

[0477] Further, the communication module of the device #1 may transmit the check result to the device #2.

[0478] Alternatively, the device #1 may send the distance(s) to the device #2. The device #2 may receive the distance(s) and check whether the encoder #1 can work with decoder #2.

[0479] Further, the communication module of the device #2 may transmit the check result to the device #1.

[0480] The above is only an example. For example, in other implementations, device # 2 can also serve as the first network element and device # 1 can also serve as the second network element. For another example, in FIG. 17, one group of reference samples is sent, and in other implementations, a plurality of groups of reference samples can be sent.

[0481] FIG. 18 is a schematic diagram of three groups of reference data samples.

[0482] For example, as shown in FIG. 18, the device #1 can be the first network element, and the device #2 can be the second network element. The relevant descriptions of the two AEs can refer to the description in FIG. 17, and will not be repeated here.

[0483] The device #2 may generate three groups of compressed reference data samples, where the group #1 of compressed reference data samples ($\mathbb{C}_1$) corresponds to the input (X_{in}) to the AE #1, the group #2 of compressed reference data samples ($\mathbb{C}_2$)corresponds to one latent layer output(X_{latent}) of the AE #1, and the group #3 of compressed reference data samples ($\mathbb{C}_3$) corresponds to the output(X_{out}) from the AE #1. Further, the AI module of the device #2 may also generate second transformation matrix #1 θ_1 and sampling matrix #1 P_1 for the group #1, second transformation matrix #2 θ_2 and sampling matrix #2 P_2 for the group #2, and second transformation matrix #3 θ_3 and sampling matrix #3 P_3 for the group #3. The communication module of the device #2 may transmit $\langle \mathbb{C}_1, \theta_1, P_1, d_1(\ ) \rangle$ for the group #1, $\langle \mathbb{C}_2, \theta_2, P_2, d_2(\ ) \rangle$ for the group #2, $\langle \mathbb{C}_3, \theta_3, P_3, d_3(\ ) \rangle$ for the group #3, with the averaged neurons to the device #1 in unicast way. $d_1(\ )$ is the scoring function for group#1, $d_2(\ )$ is the scoring function for group#2, and $d_3(\ )$ is the scoring function for group #3. The AI module of the device #1 samples and compresses the local data samples (e.g. $\hat{X}_{in}$, $\hat{X}_{latent}$ and $\hat{X}_{out}$) to obtain the three groups of the compressed local data samples $\hat{\mathbb{C}}_1$, $\hat{\mathbb{C}}_2$, and $\hat{\mathbb{C}}_3$. The AI module of the device #1

measures the distances δ_1 for the group #1, δ_2 for the group #2, and δ_3 for the group #3. After the measurement is completed, the communication module of the device #1 may transmit the distances to the device #2. Further, the communication module of the device #1 may also transmit all of the neurons or a portion of its neurons to the device #2.

[0484] The transmission process in example scenario-1 and example scenario-2 are merely examples. For other implementation methods, please refer to method 1200. For example, in FIG. 18, the communication module of the first network element may transmit a portion of three distances. For another example, the scoring function(s) may be pre-defined.

[0485] The communication method according to the embodiments of the present application is described in detail above, and the communication apparatus according to the embodiments of the present application will be described in detail below with reference to FIGS. 19-23.

[0486] FIG. 19 is a schematic block diagram of a communication apparatus 10 according to an embodiment of the present application. As shown in FIG. 19, the communication apparatus 10 includes:

[0487] a processing module 11, configured to obtain Q group(s) of first data sample(s) corresponding to Q layer(s) of an AI model, where the Q group(s) of the first data sample(s) is from compressed Q group(s) of first raw data sample(s) which is compressed according to Q transformation matrix(es), the Q group(s) of the first data sample(s) is related to an inference cycle of the AI model, and Q is a positive integer; and

[0488] a transceiver module 12, configured to send the Q group(s) of the first data sample(s).

[0489] The communication apparatus 10 in this embodiment of the present application may correspond to the second network element in the communication method in the embodiments of the present application described above, and the foregoing management operations and/or functions and other management operations and/or functions of modules of the communication apparatus 10 are intended to implement corresponding steps of the foregoing methods. For brevity, details are not described herein again.

[0490] The transceiver module 12 in this embodiment of the present application may be implemented by a transceiver, and the processing module 11 may be implemented by a processor.

[0491] As shown in FIG. 20, a communication apparatus 20 may include a transceiver 21. Optionally, the communication apparatus 20 may further include a processor 22 and/or a memory 23. The memory 23 may be configured to store indication information, or may be configured to store code, instructions, and the like that is to be executed by the processor 22.

[0492] FIG. 21 is a schematic block diagram of a communication apparatus 30 according to an embodiment of the present application. As shown in FIG. 21, the communication apparatus 30 includes:

[0493] a transceiver module 31, configured to receive Q group(s) of first data sample(s) corresponding to Q layer(s) of an AI model, where the Q group(s) of the first data sample(s) is from compressed Q group(s) of first raw data sample(s) which is compressed according to Q transformation matrix(es), the Q group(s) of the first data sample(s) is related to an inference cycle of the AI model, and Q is a positive integer.

[0494] The communication apparatus 30 in this embodiment of the present application may correspond to the first network element in the communication method in the embodiments of the present application described above, and the management operations and/or functions and other management operations and/or functions of modules of the communication apparatus 30 are intended to implement corresponding steps of the foregoing methods. For brevity, details are not described herein again.

[0495] The transceiver module 31 in this embodiment of the present application may be implemented by a transceiver.

[0496] As shown in FIG. 22, a communication apparatus 40 may include a transceiver 41. Optionally, the communication apparatus 40 may further include a processor 42 and/or a memory 43. The memory 43 may be configured to store indication information, or may be configured to store code, instructions, and the like that is to be executed by the processor 42.

[0497] The processor 22 or the processor 42 may be an integrated circuit chip and have a signal processing capability. In an embodiment process, steps in the foregoing method embodiments can be implemented by using a hardware-integrated logical circuit in the processor, or by using instructions in the form of software. The processor 22 or the processor 42 may be a general-purpose processor, a digital signal processor (DSP), an application-specific integrated circuit (ASIC), a field programmable gate array (FPGA), or another programmable logic device, a discrete gate or a transistor logic device, or a discrete hardware component. All methods, steps, and logical block diagrams disclosed in this embodiment of the present application may be implemented or performed. The general-purpose processor may be a microprocessor, or the processor may be any conventional processor or the like. Steps of the methods disclosed in the embodiments of the present invention may be directly performed and completed by a hardware decoding processor, or may be performed and completed by using a combination of hardware and software modules in the decoding processor. The software module may be located in a storage medium known in the art, such as a random access memory, a flash memory, a read-only memory, a programmable read-only memory, an electrically erasable programmable memory, or a register. The storage medium is located in the memory, and the processor reads the information in the memory and completes the steps in the foregoing methods in combination with the hardware of the processor.

[0498] It may be understood that the memory 23 or the memory 43 in the embodiments of the present invention may be a volatile memory or a non-volatile memory, or may include a volatile memory and a non-volatile memory. The non-volatile

memory may be a read-only memory (ROM), a programmable read-only memory (PROM), an erasable programmable read-only memory (EPROM), an electrically erasable programmable read-only memory EEPROM), or a flash memory. The volatile memory may be a random access memory (RAM), and be used as an external cache. Through example but not limitative description, many forms of RAMs may be used, for example, a static random access memory (SRAM), a dynamic random access memory (DRAM), a synchronous dynamic random access memory SDRAM), a double data rate synchronous dynamic random access memory (DDR SDRAM), an enhanced synchronous dynamic random access memory (Enhanced SDRAM, ESDRAM), a synchronous link dynamic random access memory (SLDRAM), and a direct rambus dynamic random access memory (DR RAM). The storage of the system and the method described in this specification aim to include, but are not limited to, these and any other proper storage.

[0499] An embodiment of the present application further provides a system. As shown in FIG. 23, a system 50 includes:

[0500] the communication apparatus 10 according to the embodiments of the present application and the communication apparatus 20 according to the embodiments of the present application.

[0501] An embodiment of the present application further provides a computer storage medium, and the computer storage medium may store one or more program instructions for executing any of the foregoing methods.

[0502] Optionally, the storage medium may be specifically the memory 23 or 43.

[0503] A person of ordinary skill in the art will be aware that, in combination with the examples described in the embodiments disclosed in this specification, units and algorithm steps may be implemented by using electronic hardware or a combination of computer software and electronic hardware. Whether the functions are performed by using hardware or software depends on particular applications and design constraint conditions of the technical solutions. A person skilled in the art may use different methods to implement the described functions for each particular application, but it should not be considered that the embodiment goes beyond the scope of the present application.

[0504] It would be understood by a person skilled in the art that, for the purpose of convenience and brevity, in a detailed working process of the foregoing system, apparatus, and unit, reference may be made to a corresponding process in the foregoing method embodiments, and details are not described herein again.

[0505] In the several embodiments provided in the present application, the disclosed system, apparatus, and method may be implemented in other manners. For example, the described apparatus embodiment is merely an example. For example, the unit division is a logical function division and other methods of division may be used in an actual embodiment. 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 using various communication interfaces. The indirect couplings or communication connections between the

apparatuses or units may be implemented in electronic, mechanical, or other forms.

[0506] 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, that is, the parts may be located in one unit, or may be distributed among a plurality of network units. Some or all of the units may be selected based on actual requirements to achieve the objectives of the embodiments.

[0507] In addition, function units in the embodiments of the present application may be integrated into one processing unit, each of the units may exist alone physically, or two or more units may be integrated into one unit.

[0508] When the functions are implemented in the form of a software functional unit and sold or used as an independent product, the functions may be stored in a computer-readable storage medium. The technical solutions of the present application may be implemented in the form of a software product. The software product is stored in a storage medium, and includes several instructions for instructing a computer device (which may be a personal computer, a server, a network device, or the like) to perform all or some of the steps of the methods described in the embodiments of the present application. The foregoing storage medium includes any medium that can store program code, such as a USB flash drive, a removable hard disk, a read-only memory (ROM), a random access memory (RAM), a magnetic disk, an optical disc or the like.

[0509] The foregoing descriptions are merely specific embodiments of the present application, but are not intended to limit the protection scope of the present application. Any variation or replacement readily figured out by a person skilled in the art within the technical scope disclosed in the present application shall fall within the protection scope of the present application. Therefore, the protection scope of the present application shall be subject to the protection scope of the claims.

CLAIMS

What is claimed is:

1. A communication method, comprising:

obtaining Q group(s) of first data sample(s) corresponding to Q layer(s) of an AI model, wherein the Q group(s) of the first data sample(s) is from compressed Q group(s) of first raw data sample(s) which is compressed according to Q transformation matrix(es), the Q group(s) of the first data sample(s) is related to an inference cycle of the AI model, and Q is a positive integer; and

sending the Q group(s) of the first data sample(s).

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

sending first information indicating the Q transformation matrix(es).

3. The communication method according to claim 2, wherein the first information is further configured to indicate Q sampling matrix(es), the Q sampling matrix(es) is configured to sample Q group(s) of second raw data sample(s), and the Q transformation matrix(es) is configured to compress sampling result(s) of the Q group(s) of the second raw data sample(s) into Q group(s) of second data sample(s).

4. The communication method according to any one of claims 1 to 3, further comprising:

receiving second information indicating difference(s) between q group(s) of second data sample(s) and q group(s) of the first data sample(s) in the Q group(s) of the first data sample(s), wherein the q group(s) of the second data sample(s) is based on inputs or outputs of q layer(s) in the Q layer(s) during the inference cycle, and q is a positive integer, $q \leqslant Q$.

5. The communication method according to claim 4, wherein the difference(s) between the q group(s) of the second data sample(s) and the q group(s) of the first data sample(s) is configured to check whether the inference cycle is abnormal.

6. The communication method according to any one of claims 1 to 5, further comprising:

sending third information indicating correspondence between the Q layer(s) and the Q group(s) of the first data sample(s).

7. The communication method according to any one of claims 1 to 6, further comprising:

sending fourth information indicating Q scoring function(s), wherein the Q scoring function(s) is configured to measure difference(s) between the Q group(s) of the first data sample(s) and Q group(s) of second data sample(s), and the Q group(s) of the second data sample(s) is based on inputs or outputs of the Q layer(s).

8. A communication method, comprising:

receiving Q group(s) of first data sample(s) corresponding to Q layer(s) of an AI model, wherein the Q group(s) of the

first data sample(s) is from compressed Q group(s) of first raw data sample(s) which is compressed according to Q transformation matrix(es), the Q group(s) of the first data sample(s) is related to an inference cycle of the AI model, and Q is a positive integer.

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

receiving first information indicating the Q transformation matrix(es).

10. The communication method according to claim 9, wherein the first information is further configured to indicate Q sampling matrix(es), the Q sampling matrix(es) is configured to sample Q group(s) of second raw data sample(s), and the Q transformation matrix(es) is configured to compress sampling result(s) of the Q group(s) of the second raw data sample(s) into Q group(s) of second data sample(s).

11. The communication method according to any one of claims 8 to 10, further comprising:

sending second information indicating difference(s) between q group(s) of second data sample(s) and q group(s) of the first data sample(s) in the Q group(s) of the first data sample(s), wherein the q group(s) of the second data sample(s) is based on inputs or outputs of q layer(s) in the Q layer(s) during the inference cycle, and q is a positive integer, $q \leqslant Q$.

12. The communication method according to claim 11, wherein the difference(s) between the q group(s) of the second data sample(s) and the q group(s) of the first data sample(s) is configured to determine whether the inference cycle of the AI model is abnormal.

13. The communication method according to any one of claims 8 to 12, further comprising:

receiving third information indicating correspondence between the Q layer(s) and the Q group(s) of the first data sample(s).

14. The communication method according to any one of claims 8 to 13, further comprising:

receiving fourth information indicating Q scoring function(s), wherein the Q scoring function(s) is configured to measure difference(s) between the Q group(s) of the first data sample(s) and Q group(s) of second data sample(s), and the Q group(s) of the second data sample(s) is based on inputs or outputs of the Q layer(s).

15. An apparatus, wherein the apparatus comprises a processor and a memory storing one or more instructions that are capable of being run on the processor, and when the one or more instructions are run, the apparatus is enabled to perform the method according to any one of claims 1 to 7 or perform the method according to any one of claims 8 to 14.

16. An apparatus, wherein the apparatus comprises a unit to perform the method according to any one of claims 1 to 7 or perform the method according to any one of claims 8 to 14.

17. A communication system, comprising a first communication apparatus and a second communication apparatus, wherein the first communication apparatus performs the method according to any one of claims 1 to 7, and the second

communication apparatus performs the method according to any one of claims 8 to 14.

18. A computer-readable storage medium, comprising one or more instructions, wherein when the one or more instructions are run on a computer, the computer performs the method according to any one of claims 1 to 7, or the method according to any one of claims 8 to 14.

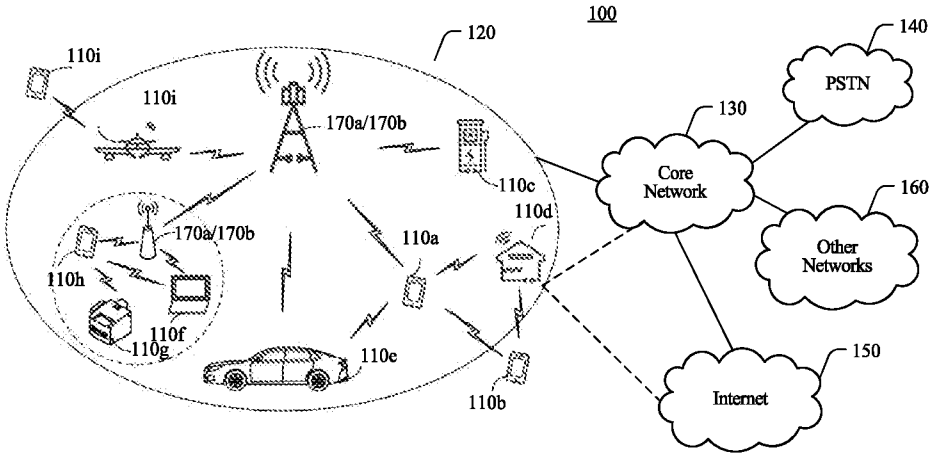
100
120
140
110i
110i
PSTN
130
170a/170b
Core Network
110c
160
110d
170a/170b
110a
Other Networks
110h
110f
110g
110e
150
Internet
110b

FIG. 1

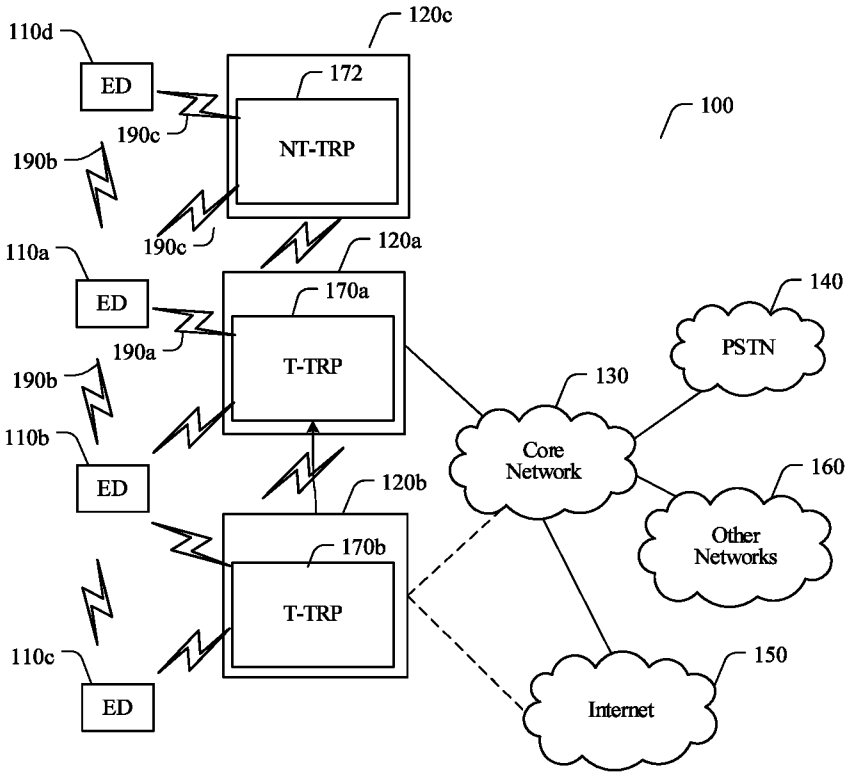
110d
120c
ED
172
100
NT-TRP
190c
190b
190c
110a
120a
ED
170a
140
T-TRP
PSTN
190a
190b
130
110b
Core Network
ED
160
120b
Other Networks
170b
T-TRP
110c
150
ED
Internet

FIG. 2

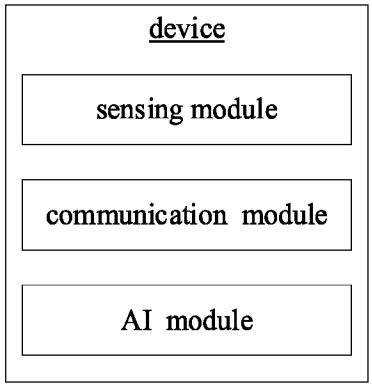

FIG. 3

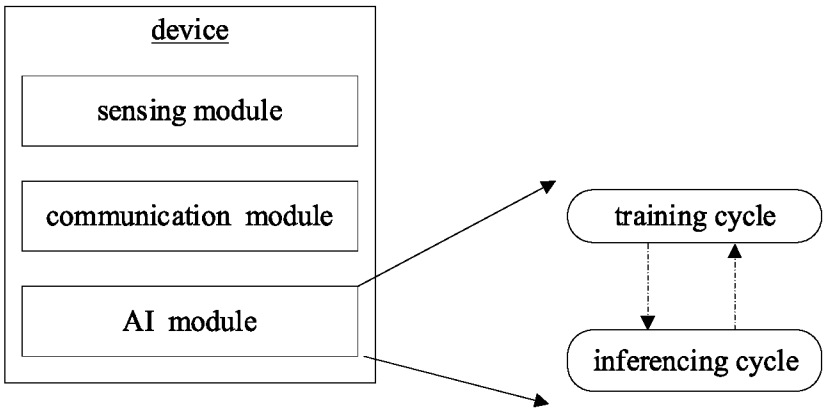

FIG. 4

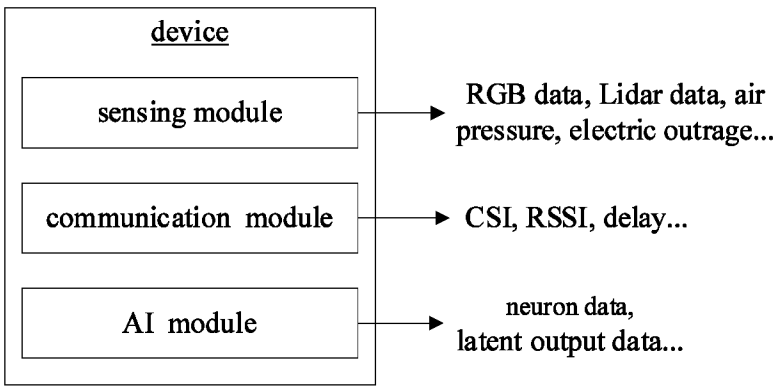

FIG. 5

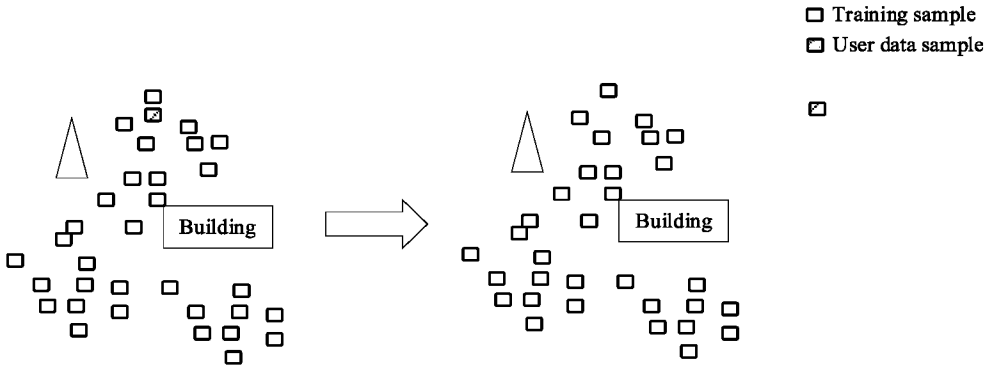

FIG. 6

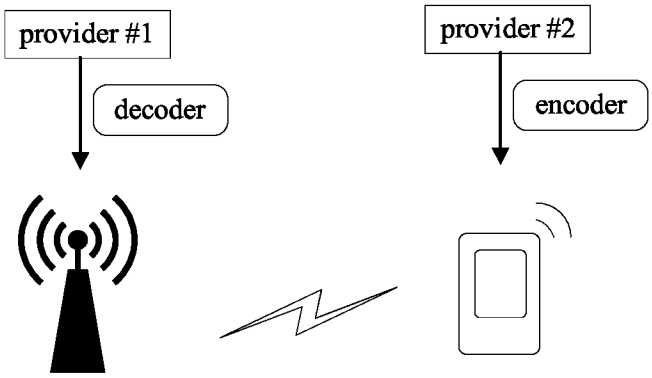

FIG. 7

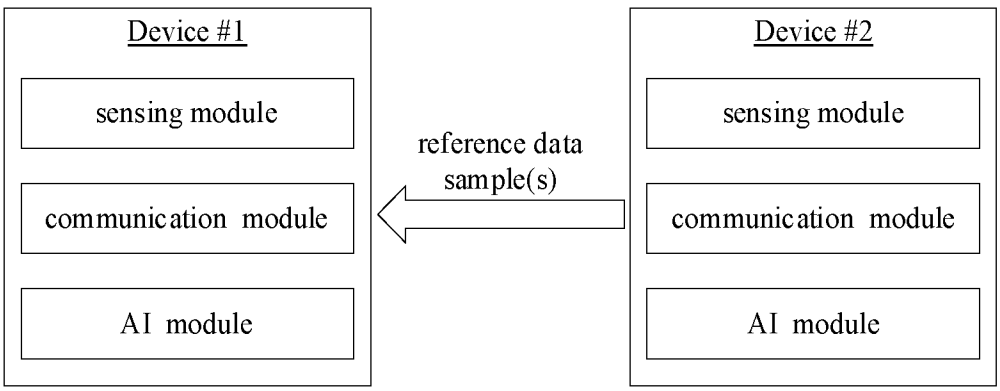

FIG. 8

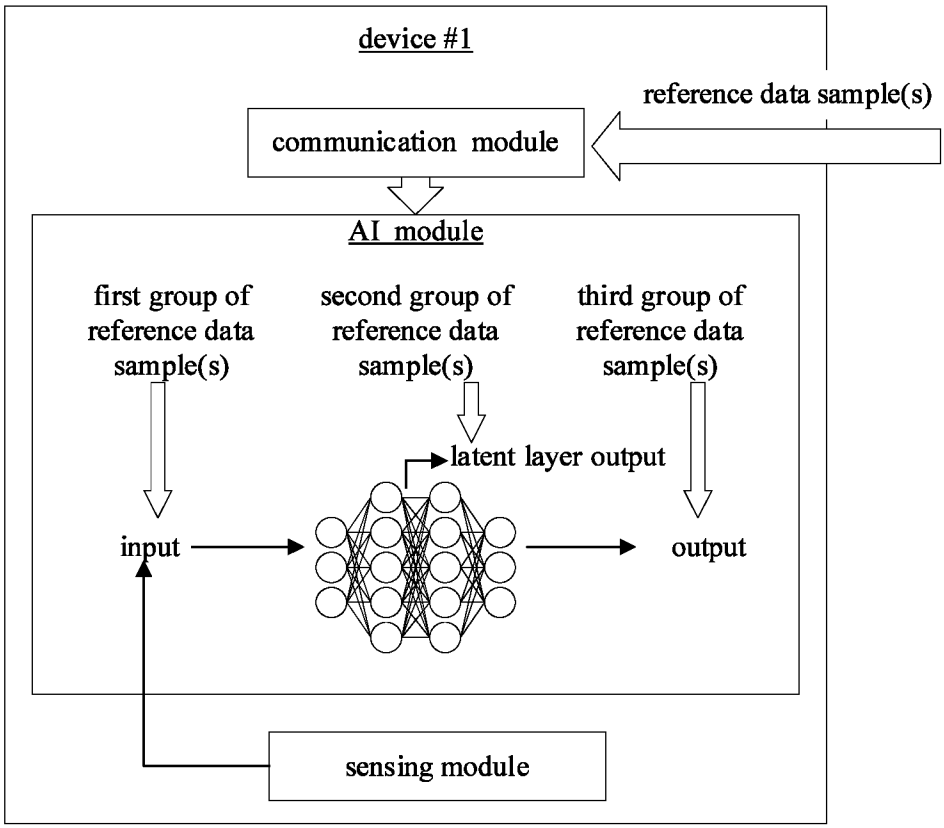
device #1
reference data sample(s)
communication module
AI module
first group of reference data sample(s)
second group of reference data sample(s)
third group of reference data sample(s)
latent layer output
input
output
sensing module

FIG. 9

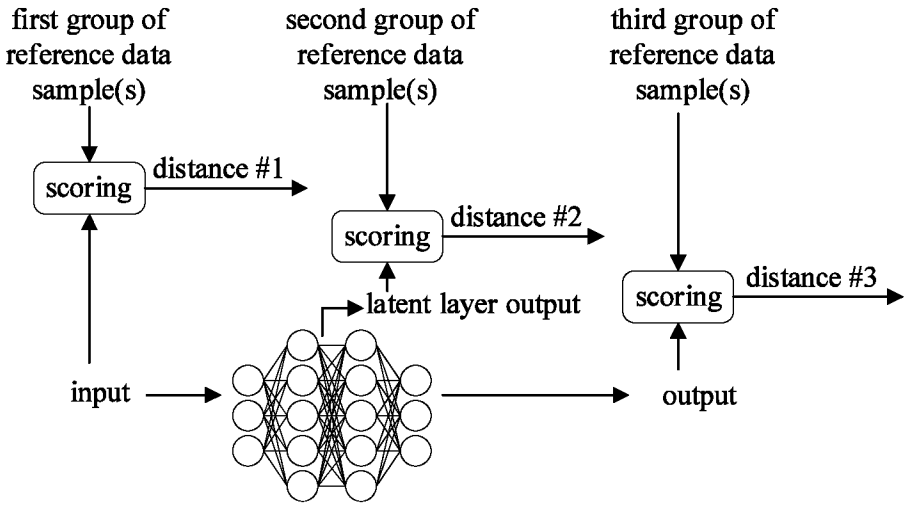
first group of reference data sample(s)
second group of reference data sample(s)
third group of reference data sample(s)
scoring
distance #1
scoring
distance #2
scoring
distance #3
latent layer output
input
output

FIG. 10

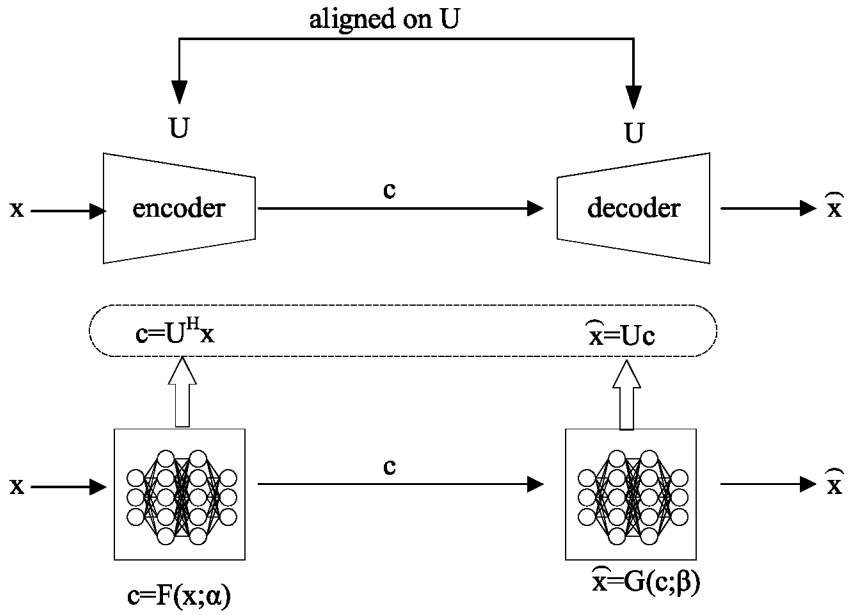
aligned on U
U
U
x
encoder
c
decoder
$\hat{x}$
$c=U^H x$
$\hat{x}=Uc$
x
c
$\hat{x}$
$c=F(x;\alpha)$
$\hat{x}=G(c;\beta)$

FIG. 11

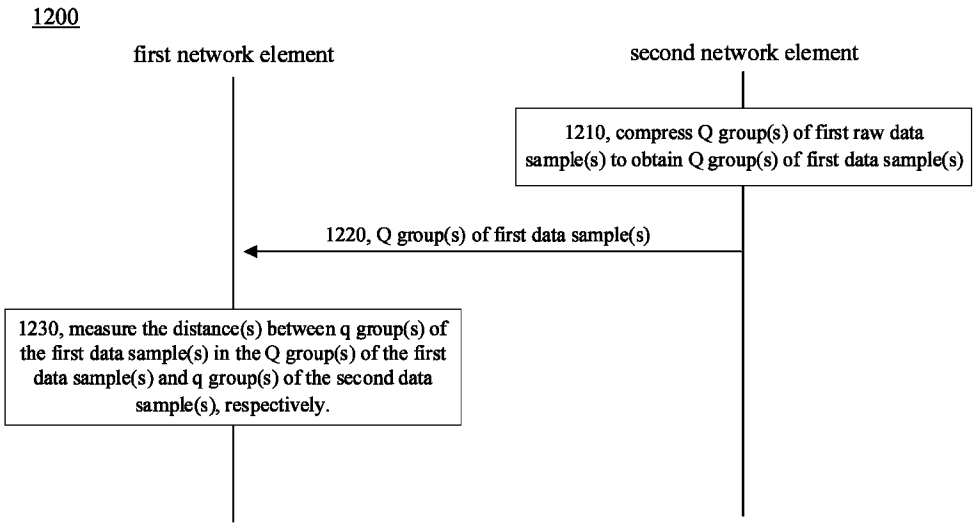
1200
first network element
second network element
1210, compress Q group(s) of first raw data sample(s) to obtain Q group(s) of first data sample(s)
1220, Q group(s) of first data sample(s)
1230, measure the distance(s) between q group(s) of the first data sample(s) in the Q group(s) of the first data sample(s) and q group(s) of the second data sample(s), respectively.

FIG. 12

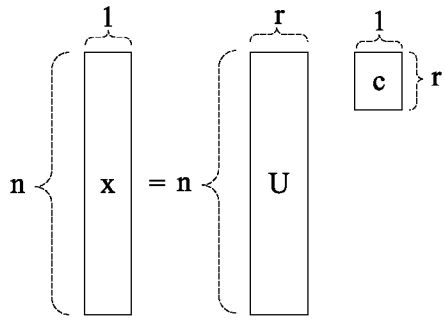

FIG. 13

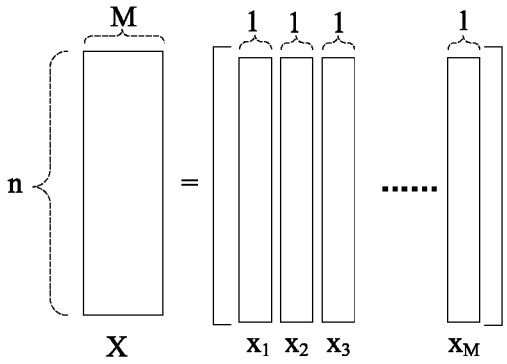

FIG. 14

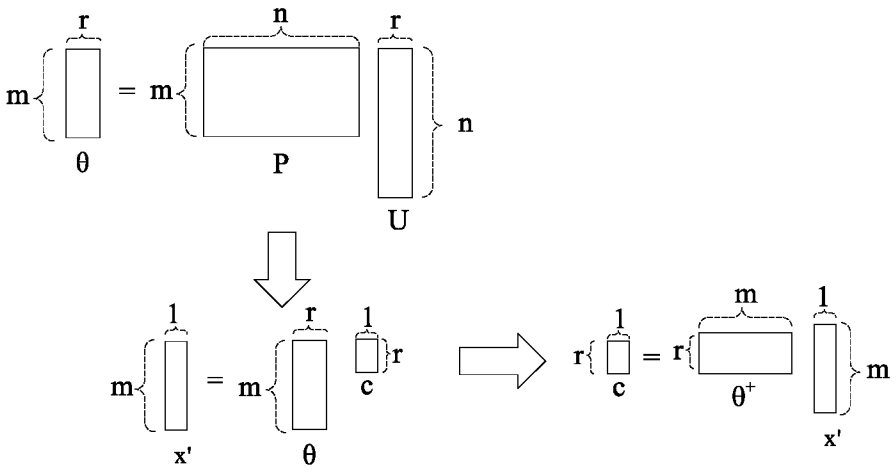
r
n
r
m
= m
n
θ
P
U
1
r
1
r
m
1
m
= m
r
c
r
1
r
= r
m
1
x'
θ
c
θ+
x'
m

FIG. 15

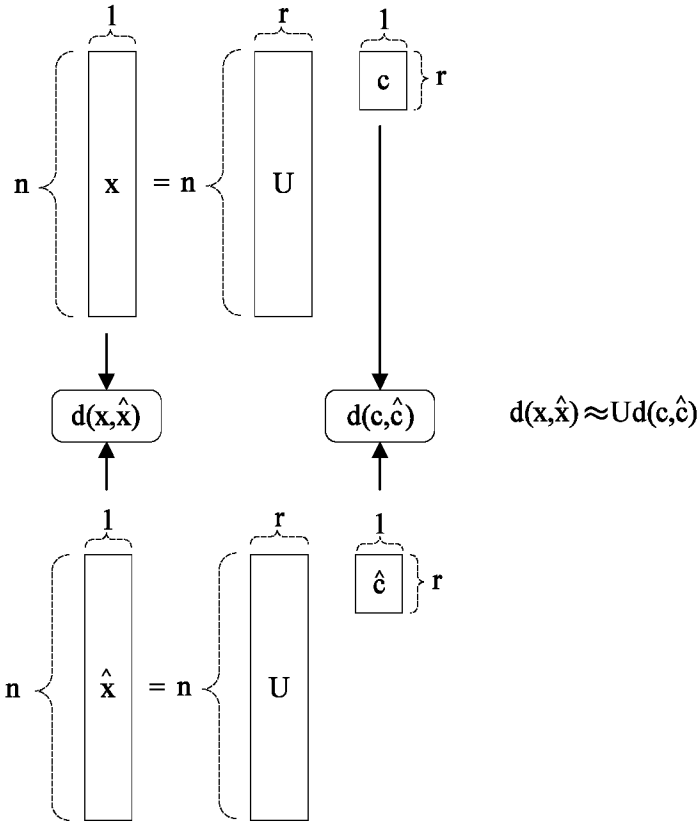
1
r
1
c
r
n
x
= n
U
d(x,x̂)
d(c,ĉ)
d(x,x̂)≈Ud(c,ĉ)
1
r
1
ĉ
r
n
x̂
= n
U

FIG. 16

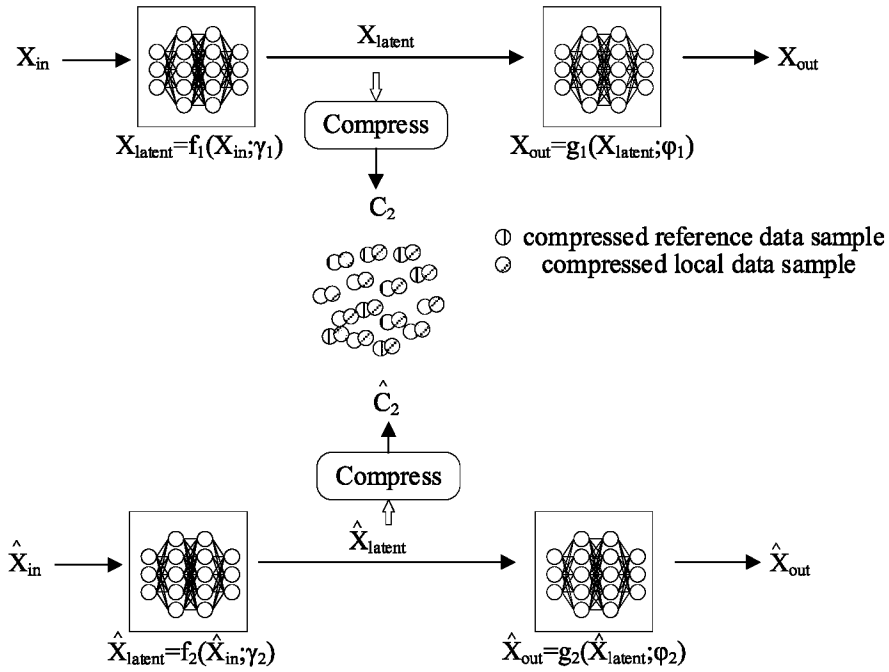
X_{in}
X_{latent}
X_{out}
Compress
$X_{latent}=f_1(X_{in};\gamma_1)$
$X_{out}=g_1(X_{latent};\varphi_1)$
C_2
compressed reference data sample
compressed local data sample
$\hat{C}_2$
Compress
$\hat{X}_{in}$
$\hat{X}_{latent}$
$\hat{X}_{out}$
$\hat{X}_{latent}=f_2(\hat{X}_{in};\gamma_2)$
$\hat{X}_{out}=g_2(\hat{X}_{latent};\varphi_2)$

FIG. 17

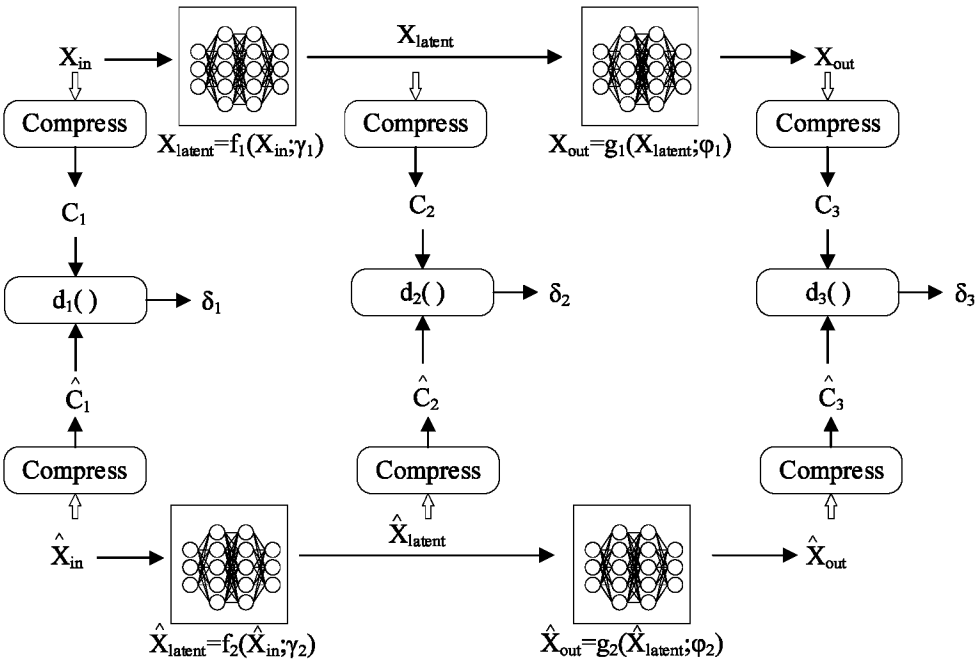
X_{in}
X_{latent}
X_{out}
Compress
$X_{latent}=f_1(X_{in};\gamma_1)$
$X_{out}=g_1(X_{latent};\varphi_1)$
C_1
C_2
C_3
$d_1(\,)$
$d_2(\,)$
$d_3(\,)$
δ_1
δ_2
δ_3
$\hat{C}_1$
$\hat{C}_2$
$\hat{C}_3$
$\hat{X}_{in}$
$\hat{X}_{latent}$
$\hat{X}_{out}$
$\hat{X}_{latent}=f_2(\hat{X}_{in};\gamma_2)$
$\hat{X}_{out}=g_2(\hat{X}_{latent};\varphi_2)$

FIG. 18

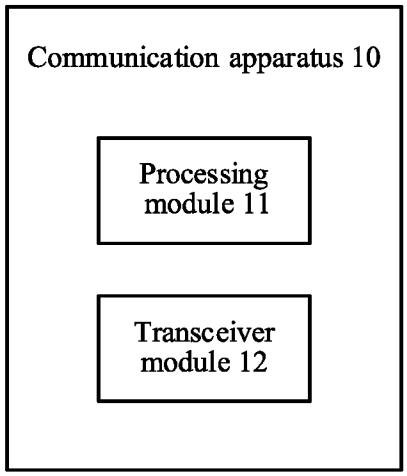
Communication apparatus 10
Processing module 11
Transceiver module 12

FIG. 19

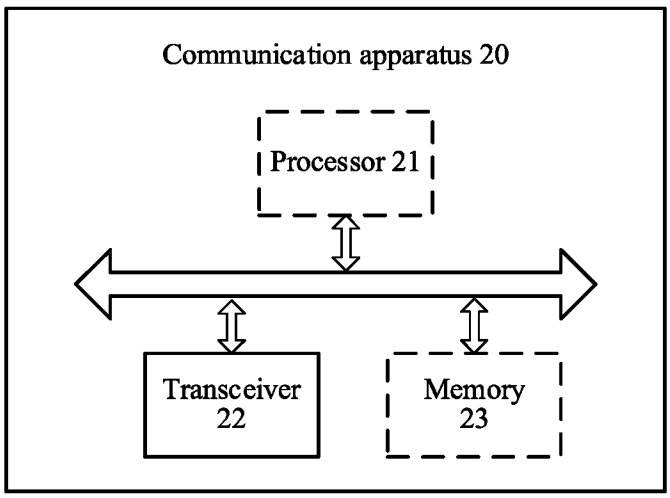
Communication apparatus 20
Processor 21
Transceiver 22
Memory 23

FIG. 20

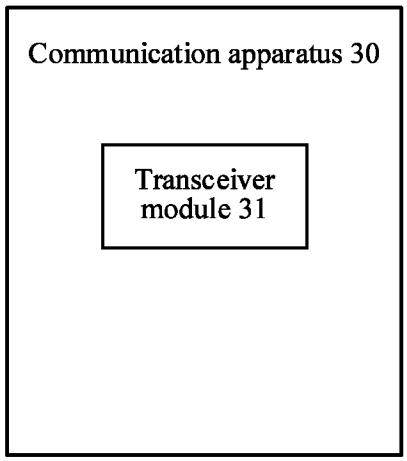

FIG. 21

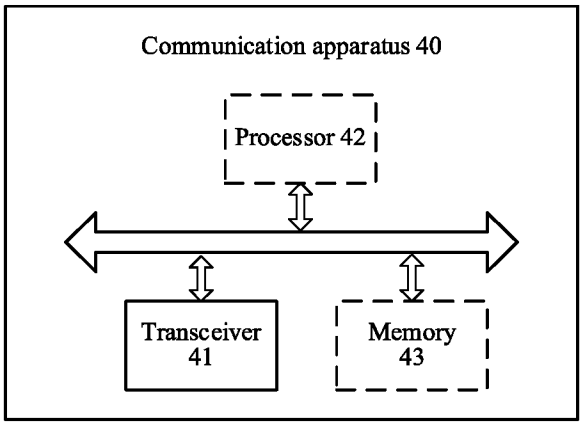
Communication apparatus 40
Processor 42
Transceiver 41
Memory 43

FIG. 22

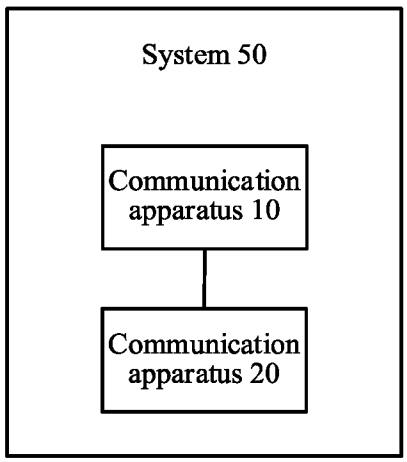
System 50
Communication apparatus 10
Communication apparatus 20

FIG. 23

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2023/125044

A. CLASSIFICATION OF SUBJECT MATTER

H04L69/04(2022.01)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:H04W,H04L,H04Q

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)

CNTXT,ENTXT,DWPI,CNKI, IEEE: AI, ML, Artificial Intelligence, model, layer, transformation, matrix, inference, cycle, compress, code

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2021279636 A1 (INTERNATIONAL BUSINESS MACHINES CORPPRATION) 09 September 2021 (2021-09-09) description, paragraphs [0036]-[0076]	1-18
A	WO 2022167547 A1 (INTERDIGITAL CE PATENT HOLDINGS) 11 August 2022 (2022-08-11) the whole document	1-18
A	CN 114357519 A (ALIPAY HANGZHOU INFORMATION TECHNOLOGY) 15 April 2022 (2022-04-15) the whole document	1-18
A	CN 114630207 A (ZHEJIANG UNIVERSITY) 14 June 2022 (2022-06-14) the whole document	1-18
A	US 2023074979 A1 (QUALCOMM INCORPORATED) 09 March 2023 (2023-03-09) the whole document	1-18

☐ 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
31 January 2024	**08 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**	**HUANG,XinXin** Telephone No. **(+86) 010-53961641**

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2023/125044

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
US	2021279636	A1	09 September 2021	WO	2021176282	A1	10 September 2021
				KR	20220133914	A	05 October 2022
				CN	115244587	A	25 October 2022
				GB	2609126	A	25 January 2023
				JP	2023516120	W	18 April 2023
				CA	3165134	A1	10 September 2021
				AU	2021231419	A1	25 August 2022
WO	2022167547	A1	11 August 2022	EP	4288907	A1	13 December 2023
				CN	116940946	A	24 October 2023
CN	114357519	A	15 April 2022	None			
CN	114630207	A	14 June 2022	None			
US	2023074979	A1	09 March 2023	TW	202312031	A	16 March 2023
				WO	2023028411	A1	02 March 2023