(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 2024/254864 A1

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

(21) International Application Number:
PCT/CN2023/100799

(22) International Filing Date:
16 June 2023 (16.06.2023)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant: **NEC CORPORATION** [JP/JP]; 7-1, Shiba 5-Chome, Minato-Ku, Tokyo 108-8001 (JP).

(71) Applicant *(for SC only)*: **WANG, Gang** [CN/CN]; 6F, Building D2, Liangmaqiao Diplomatic Office Building, No.19 Dongfangdonglu, Chaoyang District, Beijing 100600 (CN).

(72) Inventors: **GUAN, Peng**; 6F, Building D2, Liangmaqiao Diplomatic Office Building, No.19 Dongfangdonglu, Chaoyang District, Beijing 100600 (CN). **HE, Zhen**; 6F, Building D2, Liangmaqiao Diplomatic Office Building, No.19 Dongfangdonglu, Chaoyang District, Beijing 100600 (CN). **WANG, Gang**; 6F, Building D2, Liangmaqiao Diplomatic Office Building, No.19 Dongfangdonglu, Chaoyang District, Beijing 100600 (CN).

(74) Agent: **SHIHUI PARTNERS**; 42/F, Tower C, Beijing Yintai Centre, No.2 Jianguomenwai Avenue, Chaoyang District, Beijing 100022 (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,

(54) Title: DEVICES AND METHODS FOR COMMUNICATION

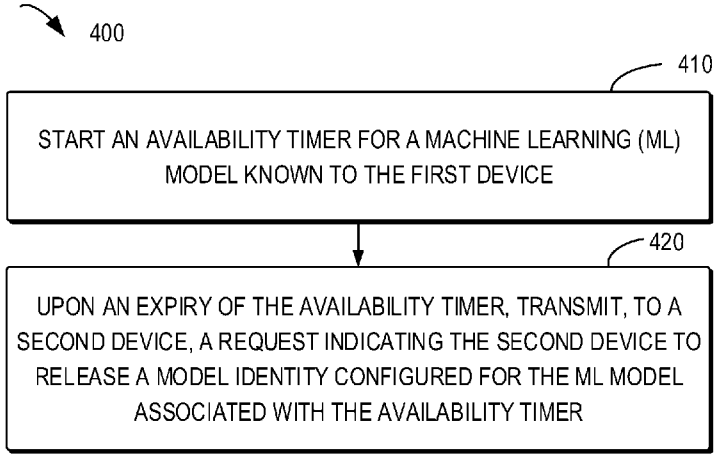

FIG. 4

(57) Abstract: Embodiments of the present disclosure provide a solution for timer-based cycle management (LCM) of machine learning (ML) model is proposed. In this solution, the first device (such as, a terminal device) starts an availability timer for a machine learning (ML) model known to the first device. Then, upon an expiry of the availability timer, first device transmits, to a second device (such as, a network device/ another terminal device), a request indicating the second device to release a model identity configured for the ML model associated with the availability timer.

WO 2024/254864 A1

[Continued on next page]

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))*

DEVICES AND METHODS FOR COMMUNICATION

FIELDS

[0001] Example embodiments of the present disclosure generally relate to the field of communication techniques and in particular, to devices and methods for timer-based life cycle management (LCM) of machine learning (ML) model.

BACKGROUND

[0002] As communication networks and services increase in size, complexity, and number of users, operations in the communication networks may become increasingly more complicated. In order to improve the communication performance, ML/artificial intelligence (AI) technology is proposed to be used in the wireless communication network. LCM is one of the core parts for AI/ML related specification. LCM may include the following stages/phases/procedure: data collection, model training, model registration, model deployment, model configuration, model inference, model selection, model activation, model deactivation, model switching, and fallback operation, model monitoring, model update, model transfer, UE capability reporting and so on. If the model cannot be identified and released properly, the resource utilization may be decreased accordingly.

SUMMARY

[0003] In general, embodiments of the present disclosure provide a solution for timer-based LCM of ML model.

[0004] In a first aspect, there is provided a first device comprising: a processor configured to cause the first device to: start an availability timer for a machine learning (ML) model known to the first device; and upon an expiry of the availability timer, transmit, to a second device, a request indicating the second device to release a model identity configured for the ML model associated with the availability timer.

[0005] In a second aspect, there is provided a second device comprising: a processor

configured to cause the second device to: receive, from a first device, a request indicating the second device to release a model identity configured for a ML model known to the first device; and release the model identity based on the request.

[0006] In a third aspect, there is provided a communication method performed by a first device. The method comprises: starting an availability timer for a machine learning (ML) model known to the first device; and upon an expiry of the availability timer, transmitting, to a second device, a request indicating the second device to release a model identity configured for the ML model associated with the availability timer.

[0007] In a fourth aspect, there is provided a communication method performed by a second device. The method comprises: receiving, from a first device, a request indicating the second device to release a model identity configured for a ML model known to the first device; and releasing the model identity based on the request.

[0008] In a fifth aspect, there is provided a computer readable medium having instructions stored thereon, the instructions, when executed on at least one processor, causing the at least one processor to carry out the method according to the third, or fourth aspect.

[0009] Other features of the present disclosure will become easily comprehensible through the following description.

BRIEF DESCRIPTION OF THE DRAWINGS

[0010] Through the more detailed description of some example embodiments of the present disclosure in the accompanying drawings, the above and other objects, features and advantages of the present disclosure will become more apparent, wherein:

[0011] FIG. 1A illustrates an example communication environment in which example embodiments of the present disclosure can be implemented;

[0012] FIG. 1B illustrates an example framework 100B of a ML model;

[0013] FIG. 2A illustrates a signaling flow for communication in accordance with some embodiments of the present disclosure;

[0014] FIG. 2B illustrates a hierarchical structure of global model identity;

[0015] FIG. 2C illustrates a concatenation structure of global model ID;

[0016] FIG. 3A and 3C illustrate signaling flows for identification procedures in accordance with some embodiments of the present disclosure;

[0017] FIG. 4 illustrates a flowchart of a method implemented at a first device according to some example embodiments of the present disclosure;

[0018] FIG. 5 illustrates a flowchart of a method implemented at a second device according to some example embodiments of the present disclosure;

[0019] FIG. 6 illustrates a simplified block diagram of an apparatus that is suitable for implementing example embodiments of the present disclosure.

[0020] Throughout the drawings, the same or similar reference numerals represent the same or similar element.

DETAILED DESCRIPTION

[0021] Principle of the present disclosure will now be described with reference to some example embodiments. It is to be understood that these embodiments are described only for the purpose of illustration and help those skilled in the art to understand and implement the present disclosure, without suggesting any limitation as to the scope of the disclosure. Embodiments described herein can be implemented in various manners other than the ones described below.

[0022] In the following description and claims, unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skills in the art to which this disclosure belongs.

[0023] As used herein, the term 'terminal device' refers to any device having wireless or wired communication capabilities. Examples of the terminal device include, but not limited to, user equipment (UE), personal computers, desktops, mobile phones, cellular phones, smart phones, personal digital assistants (PDAs), portable computers, tablets, wearable devices, internet of things (IoT) devices, Ultra-reliable and Low Latency Communications (URLLC) devices, Internet of Everything (IoE) devices, machine type communication (MTC) devices, devices on vehicle for V2X communication where X means pedestrian, vehicle, or infrastructure/network, devices for Integrated Access and

Backhaul (IAB), Space borne vehicles or Air borne vehicles in Non-terrestrial networks (NTN) including Satellites and High Altitude Platforms (HAPs) encompassing Unmanned Aircraft Systems (UAS), eXtended Reality (XR) devices including different types of realities such as Augmented Reality (AR), Mixed Reality (MR) and Virtual Reality (VR), the unmanned aerial vehicle (UAV) commonly known as a drone which is an aircraft without any human pilot, devices on high speed train (HST), or image capture devices such as digital cameras, sensors, gaming devices, music storage and playback appliances, or Internet appliances enabling wireless or wired Internet access and browsing and the like. The 'terminal device' can further has 'multicast/broadcast' feature, to support public safety and mission critical, V2X applications, transparent IPv4/IPv6 multicast delivery, IPTV, smart TV, radio services, software delivery over wireless, group communications and IoT applications. It may also incorporate one or multiple Subscriber Identity Module (SIM) as known as Multi-SIM. The term "terminal device" can be used interchangeably with a UE, a mobile station, a subscriber station, a mobile terminal, a user terminal or a wireless device.

[0024] The term "network device" refers to a device which is capable of providing or hosting a cell or coverage where terminal devices can communicate. Examples of a network device include, but not limited to, a Node B (NodeB or NB), an evolved NodeB (eNodeB or eNB), a next generation NodeB (gNB), a transmission reception point (TRP), a remote radio unit (RRU), a radio head (RH), a remote radio head (RRH), an IAB node, a low power node such as a femto node, a pico node, a reconfigurable intelligent surface (RIS), and the like. Further, the network device also may be an operation administration and maintenance (OAM), a server, location management function (LMF), access and mobility management function (AMF), and other core network node.

[0025] The terminal device or the network device may have Artificial intelligence (AI) or Machine learning capability. It generally includes a model which has been trained from numerous collected data for a specific function, and can be used to predict some information.

[0026] The terminal device or the network device may work on several frequency ranges, e.g., FR1 (e.g., 450 MHz to 6000 MHz), FR2 (e.g., 24.25GHz to 52.6GHz), frequency band larger than 100A GHz as well as Tera Hertz (THz). It can further work on licensed/unlicensed/shared spectrum. The terminal device may have more than one connection with the network devices under Multi-Radio Dual Connectivity (MR-DC) application scenario.

The terminal device or the network device can work on full duplex, flexible duplex and cross division duplex modes.

[0027] The embodiments of the present disclosure may be performed in test equipment, e.g., signal generator, signal analyzer, spectrum analyzer, network analyzer, test terminal device, test network device, channel emulator. In some embodiments, the terminal device may be connected with a first network device and a second network device. One of the first network device and the second network device may be a master node and the other one may be a secondary node. The first network device and the second network device may use different radio access technologies (RATs). In some embodiments, the first network device may be a first RAT device and the second network device may be a second RAT device. In some embodiments, the first RAT device is eNB and the second RAT device is gNB. Information related with different RATs may be transmitted to the terminal device from at least one of the first network device or the second network device. In some embodiments, first information may be transmitted to the terminal device from the first network device and second information may be transmitted to the terminal device from the second network device directly or via the first network device. In some embodiments, information related with configuration for the terminal device configured by the second network device may be transmitted from the second network device via the first network device. Information related with reconfiguration for the terminal device configured by the second network device may be transmitted to the terminal device from the second network device directly or via the first network device.

[0028] As used herein, the singular forms 'a', 'an' and 'the' are intended to include the plural forms as well, unless the context clearly indicates otherwise. The term 'includes' and its variants are to be read as open terms that mean 'includes, but is not limited to.' The term 'based on' is to be read as 'at least in part based on.' The term 'one embodiment' and 'an embodiment' are to be read as 'at least one embodiment.' The term 'another embodiment' is to be read as 'at least one other embodiment.' The terms 'first,' 'second,' and the like may refer to different or same objects. Other definitions, explicit and implicit, may be included below.

[0029] In some examples, values, procedures, or apparatus are referred to as 'best,' 'lowest,' 'highest,' 'minimum,' 'maximum,' or the like. It will be appreciated that such descriptions are intended to indicate that a selection among many used functional alternatives can be made, and such selections need not be better, smaller, higher, or otherwise preferable to other selections.

[0030] As used herein, the term "resource," "transmission resource," "uplink resource,"

or "downlink resource" may refer to any resource for performing a communication, such as a resource in time domain, a resource in frequency domain, a resource in space domain, a resource in code domain, or any other resource enabling a communication, and the like. In the following, unless explicitly stated, a resource in both frequency domain and time domain will be used as an example of a transmission resource for describing some example embodiments of the present disclosure. It is noted that example embodiments of the present disclosure are equally applicable to other resources in other domains.

[0031] As used herein, an AI/ML model may be equivalent to at least one of the following: a model, an ML model, an AI model, a data-driven, a data processing model, an algorithm, a functionality, a procedure, a process, an entity, a function, a feature, a feature group, a model ID, a functionality ID, a configuration ID, a scenario ID, a site ID, or a dataset ID. As a result, the above terms may be used interchangeably.

[0032] In some embodiments, the AI/ML model may be represented by or associated with a channel, a resource, a resource set, a RS resource, a RS resource set, a RS port, a set of RS ports, a RS port ID, or a set of RS port IDs.

[0033] In some embodiments, the AI/ML model may comprise a set of weights values that may be learned during training, for example for a specific architecture or configuration, where a set of weights values may also be called a parameter set.

[0034] In some embodiments, the AI/ML model may be used to predict a target cell, or measurements of a set of beams of a set of candidate cells in future based on at least historical measurements (e.g., L1-RSRP, L1-SINR) of a set of beams of a set of candidate cells.

[0035] In some embodiments, an input of the AI/ML model (i.e., AI input) may refer to the input of a model and indicate data inputted into the model, which may be equivalent to data, input data or input data item.

[0036] In some embodiments, an output of AI/ML model (i.e., AI output) may refers to the output of a model and indicate result(s) outputted by the model, which is equivalent to label, data, output data, or label data.

[0037] The terms "a compression ratio (CR)" may refer to one of the following definitions (note, other Definitions are not excluded):

- Definition 1: CR = output dimension of encoder/input dimension of encoder. For

example, the input dimension of encoder may refer to the number of ports × the number of sub-bands, and the output dimension of encoder may refer to the number of compressed bits or the number of compressed bits / the number of quantization bits;

- Definition 2: CR = 2 × output dimension of encoder / input dimension of encoder. 2 may stand for real and imaginary part of a complex value. 2 may stand for dual polarizations. 2 may also be replaced by other constants;
- Definition 3: CR = the number of quantization bits × output dimension of encoder / input dimension of encoder;
- Definition 4: CR = 2 × the number of quantization bits × output dimension of encoder / input dimension of encoder;
- Definition 5: CR = the number of bits required to report codebook-based CSI / the number of bits required to report compressed CSI. The codebook-based CSI may refer to type I single panel codebook, type I multi-panel codebook, type II codebook, type II port selection codebook, enhanced type II codebook, enhanced type II port selection codebook, and further enhanced type II port selection codebook.

[0038] The expression "a higher compression ratio" in the present disclosure means less bits to compress the same amount of information, i.e., CR1 = ¼ is a higher comparison ratio compared to CR2 = ½. But in terms of the value of the fractional number, CR1<CR2. In other words, "CR1<CR2" means "CRI is a higher compression ratio than CR2" and "with CR1, less bits are required to compress the same amount of information".

[0039] In some embodiments, CR may be not a specific value of CR and may be an index value indicating CR. For example, CR index 1 may correspond to 1/2, CR index 2 may correspond to 1/4, etc.;

[0040] In some embodiments, CR may comprise no compression. For example, CR index 0 may correspond to "no compression".

[0041] The terms "a compression (procedure)" may be associated with at least one of the following parameters: a CR, the number of compressed bits (or the payload size of compressed CSI), the number of quantization bits, output or input dimension of encoder, output or input dimension of decoder, the number of layers in an encoder or decoder, dimension or size of each layer, activation function, and so on. Each of these parameters may be set per frequency unit, per time unit or per spatial unit. For example, for a CR, a

CR per frequency unit refers to the number of compressed bits per sub-band, per PRB, etc.. A CR per time unit refers to the number of compressed bits per slot, per frame, etc.. A CR per spatial unit refers to the number of compressed bits per port, per beam, etc..

[0042] The terms "quantization information" may refer to at least one of the following:

- The number of quantization bits, such as 1-bit, 2-bit, 4-bit, etc.;
- The method of quantization, such as uniform quantization, non-uniform quantization, vector quantization, scale quantization;
- A ceil, floor or round operation applied in quantization function, such as ceil/floor/round value of $(x \times 2^{(B-1)})/(2^{(B-1)})$ if B-bit uniform quantization is used;
- If the total number of compressed bits is fixed, the higher compression ratio means more bits for quantization;
- The quantization information may be not a specific value and may be an index value indicating the quantization information. For example, quantization index 0 may correspond to 1-bit quantization, quantization index 1 may correspond to 2-bit quantization, etc..

[0043] As discussed above, LCM is one of the core parts for AI/ML related specification. Currently, two types of LCM are proposed to be supported, i.e., functionality based LCM and model-identity (ID) based LCM. As a result, functionality/model identification may be needed accordingly. Further, it has been agreed that both the network device/terminal device may initiate the identification procedure.

[0044] In some embodiments, during LCM procedure, an AI/ML model may have a model ID with associated information and/or model functionality at least for some AI/ML operations. Specifically, a model may be identified by a model ID. Further, a model ID may be used to identify which AI/ML model is being used in LCM including model delivery, and the model ID may be used to identify a model (or models) during model selection/activation/deactivation/switching.

[0045] In some embodiments, for UE-part/UE-side models, in case of functionality-based LCM procedure, indication of activation/deactivation/switching/fallback may be implemented based on individual AI/ML functionality.

[0046] In some embodiments, for UE-part/UE-side models, in case of model-ID-based LCM procedure, indication of model selection/activation/deactivation/switching/fallback may be implemented based on individual model IDs.

[0047] In some embodiments, the UE may have one AI/ML model for the functionality, Alternatively, in some embodiments, the UE may have multiple AI/ML models for the functionality.

[0048] In some embodiments, for UE-side models and UE-part of two-sided models: 1) for AI/ML functionality identification legacy 3GPP framework of features may be reused, UE may indicate the supported functionalities/functionality for a given sub-use-case, and the UE capability reporting may be taken as starting point; 2) for AI/ML model identification, models may be identified by model ID at the Network and UE may indicate the supported AI/ML models; 3) in functionality-based LCM, network may indicate activation/deactivation/fallback/switching of AI/ML functionality via a signaling (e.g., a radio resource control (RRC) signalling, downlink control information (DCI)); models may not be identified at the Network, and UE may perform model-level LCM; 4) in model-ID-based LCM, models are identified at the network, and network/UE may activate/deactivate/select/switch individual AI/ML models via model ID.

[0049] In some embodiments, for functionality identification, there may be either one or more than one functionality defined within an AI/ML-enabled feature.

[0050] In some embodiments, for AI/ML model identification and model-ID-based LCM of UE-side models and/or UE-part of two-sided models, the model-ID-based LCM may operate based on identified models, where a model may be associated with specific configurations/conditions associated with UE capability of an AI/ML-enabled feature/feature group and additional conditions (e.g., scenarios, sites, and datasets) as determined/identified between UE-side and network-side.

[0051] In some embodiments, for model identification of UE-side or UE-part of two-sided models, model identification types may be categorized as follows:

- Type A: Model is identified to network (if applicable) and UE (if applicable) without over-the-air signaling, where the model may be assigned with a model ID during the model identification, which may be referred/used in over-the-air signaling after model identification;

- Type B: Model is identified via over-the-air signaling,
  - Type B1: model identification initiated by the UE, and network assists the remaining steps (if any) of the model identification and the model may be assigned with a model ID during the model identification;
  - Type B2: model identification initiated by the network, and UE responds (if applicable) for the remaining steps (if any) of the model identification, and the model may be assigned with a model ID during the model identification.

[0052] In some embodiments, for functionality/model-ID based LCM, once functionalities/models are identified, the same or similar procedures may be used for their activation, deactivation, switching, fallback, and monitoring.

[0053] In some embodiments, once models are identified, UE may indicate the supported AI/ML model IDs for a given AI/ML-enabled feature/feature group in a UE capability report as starting point.

[0054] Generally speaking, model ID should be unique "globally", e.g., in order to manage test certification, each retrained version needs to be identified. However, it is still pending whether a global model ID is needed for model inference in model activation/deactivation/selection/switching and so on since for these purposes it may not be efficient to use global model ID. In this event, it is expected that a local (or temporary) model ID may be configured for the model, which may be used for model management after identification/registration.

[0055] If the concept of a local (or temporary) model ID is introduced, how to maintain the model ID in the LCM is desirable to be further discussed, because if the local (or temporary) model ID cannot be released in time, the unnecessary signaling overhead may be increased and also causes unnecessary consumption of resources (such as, UE memory/battery) used for maintaining models which are not in use.

[0056] According to the example embodiments of the present discourse, a solution for timer-based LCM of ML model is proposed. In this solution, the first device (such as, a terminal device) starts an availability timer for a ML model known to the first device. Then, upon an expiry of the availability timer, first device transmits, to a second device (such as, a network device/another terminal device), a request indicating the second device to release a model identity configured for the ML model associated with the availability

timer.

[0057] In this way, when the LCM of a ML model known to the first device is not active (or the ML model is not in use), another device may release the model identity configured for the ML model in time. As a result, a recycling rate pf the available model identity is increased. Further, due to the timely release, the resources (such as, UE memory, or battery) used for maintaining models which are not in use may be saved accordingly.

[0058] Further, for better understanding, some terminologies related with ML model are described in below table 1.

Table 1 descriptions of terminology

Terminology	Description
AI/ML Model	A data driven algorithm that applies AI/ML techniques to generate a set of outputs based on a set of inputs.
AI/ML model delivery	A generic term referring to delivery of an AI/ML model from one entity to another entity in any manner. Note: An entity could mean a network node/function (e.g., gNB, LMF, etc.), UE, proprietary server, etc.
AI/ML model Inference	A process of using a trained AI/ML model to produce a set of outputs based on a set of inputs.
AI/ML model testing	A subprocess of training, to evaluate the performance of a final AI/ML model using a dataset different from one used for model training and validation. Differently from AI/ML model validation, testing does not assume subsequent tuning of the model.
AI/ML model training	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.
AI/ML model transfer	Delivery of an AI/ML model over the air interface in a manner that is not transparent to 3GPP signalling, 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.
AI/ML model validation	A subprocess of training, to evaluate the quality of an AI/ML model using a dataset different from one used for model training, that helps selecting model parameters that generalize beyond the dataset used for model training.
Data collection	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.
Federated learning / federated training	A machine learning technique that trains an AI/ML model across multiple decentralized edge nodes (e.g., UEs, gNBs) each performing local model training using local data samples. The technique requires multiple interactions of the model, but no exchange of local data samples.
Functionality identification	A process/method of identifying an AI/ML functionality for the common understanding between the network and the UE. Note: Information regarding the AI/ML functionality may be shared during functionality identification. Where AI/ML functionality resides depends on the specific use cases and sub use cases.
Model activation	enable an AI/ML model for a specific AI/ML-enabled feature.
Model deactivation	disable an AI/ML model for a specific AI/ML-enabled feature.
Model download	Model transfer from the network to UE.
Model identification	A process/method of identifying an AI/ML model for the common understanding between the network and the UE. Note: The process/method of model identification may or may not be applicable. Information regarding the AI/ML model may be shared during model identification.
Model monitoring	A procedure that monitors the inference performance of the AI/ML model.

Model parameter update	Process of updating the model parameters of a model.
Model selection	The process of selecting an AI/ML model for activation among multiple models for the same AI/ML enabled feature. Note: Model selection may or may not be carried out simultaneously with model activation.
Model switching	Deactivating a currently active AI/ML model and activating a different AI/ML model for a specific AI/ML-enabled feature.
Model update	Process of updating the model parameters and/or model structure of a model.
Model upload	Model transfer from UE to the network.
Network-side (AI/ML) model	An AI/ML Model whose inference is performed entirely at the network.
Offline field data	The data collected from field and used for offline training of the AI/ML model.
Offline training	An AI/ML training process where the model is trained based on collected dataset, and where the trained model is later used or delivered for inference. Note: This definition only serves as a guidance. There may be cases that may not exactly conform to this definition but could still be categorized as offline training by commonly accepted conventions.
Online field data	The data collected from field and used for online training of the AI/ML model.
Online training	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. Note: the notion of (near) real-time and non real-time are context-dependent and is relative to the inference time-scale. Note: This definition only serves as a guidance. There may be cases that may not exactly conform to this definition but could still be categorized as online training by commonly accepted conventions. Note: Fine-tuning/re-training may be done via online or offline

	training. (This note could be removed when we define the term fine-tuning.)
Reinforcement Learning (RL)	A process of training an AI/ML model from input (a.k.a. state) and a feedback signal (a.k.a. reward) resulting from the model's output (a.k.a. action) in an environment the model is interacting with.
Semi-supervised learning	A process of training a model with a mix of labelled data and unlabelled data.
Supervised learning	A process of training a model from input and its corresponding *labels*.
Two-sided (AI/ML) model	A paired AI/ML Model(s) over which joint inference is performed, where joint inference comprises AI/ML Inference whose inference is performed jointly across the UE and the network, i.e, the first part of inference is firstly performed by UE and then the remaining part is performed by gNB, or vice versa.
UE-side (AI/ML) model	An AI/ML Model whose inference is performed entirely at the UE.
Unsupervised learning	A process of training a model without labelled data.
Proprietary-format models	ML models of vendor-/device-specific proprietary format, from 3GPP perspective. They are not mutually recognizable across vendors and hide model design information from other vendors when shared. Note: An example is a device-specific binary executable format.
Open-format models	ML models of specified format that are mutually recognizable across vendors and allow interoperability, from 3GPP perspective. They are mutually recognizable between vendors and do not hide model design information from other vendors when shared.

[0059] Further, for better descriptions, some terms used herein are listed as below:

- "Conditions": referred to as configurations supported indicated via UE capability reporting (e.g., component field), related to model training, model inference, performance monitoring, validation procedure, fallback, of an AI/ML model/functionality or a group of AI/ML models/functionalities;
- "Additional conditions": including but not limited to, application conditions, scenarios, datasets, cell ID, timestamp and SNR, and so on;
- "UE internal conditions": including but not limited to, memory, battery, computation resource, overheating and other hardware limitations;
- AI/ML-enabled feature: referred to as a feature where AI/ML may be used;
- Physical AI/ML model(s): referred to as an actual implementation of such a model;
- A logical AI/ML model: referred to as a model that is identified and assigned a model ID;
- A local model index refers to a temporary model index for model control after Model identification;
- Local Model index may be allocated between network and UE and utilized for various LCM signalling purposes such as activation/deactivation/selection/switching;
- Physical model ID: used for identifying a real AI/ML model, a real implementation;
- Global model ID: usually used in radio access network (RAN) 2 solutions;
- Logical model ID: associated with one or a group of physical models for the same purpose; Local model ID: RAN1 solutions usually do not require global ID;
- "A group of models": referred to as a group of physical/logical/global/local models;
- Functionality: in case of AI/ML functionality identification and functionality-based LCM of UE-side models and/or UE-part of two-sided models, functionality is referred to an AI/ML-enabled Feature/ feature group enabled by configuration(s), where configuration(s) is(are) supported based on conditions indicated by UE capability. Correspondingly, functionality-based LCM may operate based on, at least, one configuration of AI/ML-enabled Feature/ feature group ,or specific configurations of an AI/ML-enabled Feature/ Feature group.

[0060] In the present disclosure,

- terms of "ML model", "AI model", "ML function", "AI function" and "algorithm" may be used interchangeably.
- terms of "model" "functionality" and "model/functionality" may be used interchangeably.
- terms of "model" and "model group" may be used interchangeably.
- terms of "functionality", "functionality group", "functionality set" may be used interchangeably.
- terms of "ID", "index", "indicator" and "identifier" may be used interchangeably.
- terms of "known/unknown", "applicable/non-applicable" or "suitable/non-suitable" may be used interchangeably.

[0061] In the context of the present disclosure,

- wording of "in response to/upon a model inference/model activation/switching-to/selection/model update/fine-tuning/model retraining" may refer to "in response to/upon a start, initiation or completion of a model inference/model activation/switching-to/selection/model update/fine-tuning/model retraining";
- wording of "in response to/upon a positive/negative assessment result" may refer to "in response to/upon detecting/determining a positive/negative assessment result";
- wording of "in response to/upon a message/indication" may refer to "in response to/upon receiving a message/indication" or "after a pre-configured period from receiving a message/indication".

[0062] Principles and implementations of the present disclosure will be described in detail below with reference to the figures.

Example Environment

[0063] FIG. 1A illustrates a schematic diagram of an example communication environment 100A in which example embodiments of the present disclosure can be implemented. In the communication environment 100A, a plurality of communication devices, including a first

device 110 and a second device 120, can communicate with each other.

[0064] In the example of FIG. 1A, in some embodiments, both the first device 110 and the second device 120 may include a terminal device. In this specific example embodiment, the first device 110 and the second device 120 may communicate with each other via sidelink communication.

[0065] In the example of FIG. 1A, in some embodiments, the first device 110 may include a terminal device and the second device 120 may include a network device serving the terminal device. In this specific example embodiment, a link from the first device 110 to the second device 120 is referred to as uplink, while a link from the second device 120 to the first device 110 is referred to as a downlink.

[0066] In downlink, the second device 120 is a transmitting (TX) device (or a transmitter) and the first device 110 is a receiving (RX) device (or a receiver), and the second device 120 may transmit downlink transmission to the first device 110. Correspondingly, in uplink, the second device 120 is an RX device (or a receiver) and the first device 110 is a TX device (or a transmitter), and the first device 110 may transmit uplink transmission to the second device 120.

[0067] Further, in FIG. 1A, one or more ML models may be deployed at the first device 110. Reference is now made to FIG. 1B, which illustrates an example framework 100B of a ML model.

[0068] As illustrated in FIG. 1B, the example framework 100B consists of: (i) Data Collection, (ii) Model Training, (iii) Model Management, (iv) Model Inference, and (v) Model Storage. The framework 100B may be suitable for both the model based and/or functionality based LCM. Further, "Model Storage" in the FIG. 1B is only intended as a reference point (if any) for protocol terminations etc for model transfer/delivery etc. It is not intended to limit where models are actually stored. It is to be understood that the example framework 100B should is given for illustrative purpose only, which should not be interpreted as any limitation to the present disclosure.

[0069] It is to be understood that the number of devices and their connections shown in FIG. 1A are only for the purpose of illustration without suggesting any limitation. The communication environment 100A may include any suitable number of devices configured to implementing example embodiments of the present disclosure.

[0070] In some embodiments, the first device 110 and the second device 120 may

communicate with each other via a channel such as a wireless communication channel on an air interface (e.g., Uu interface) or a PC5 interface. The wireless communication channel may comprise a sidelink, a physical uplink control channel (PUCCH), a physical uplink shared channel (PUSCH), a physical random-access channel (PRACH), a physical downlink control channel (PDCCH), a physical downlink shared channel (PDSCH) and a physical broadcast channel (PBCH). Of course, any other suitable channels are also feasible.

[0071] The communications in the communication environment 100A may conform to any suitable standards including, but not limited to, Global System for Mobile Communications (GSM), Long Term Evolution (LTE), LTE-Evolution, LTE-Advanced (LTE-A), New Radio (NR), Wideband Code Division Multiple Access (WCDMA), Code Division Multiple Access (CDMA), GSM EDGE Radio Access Network (GERAN), Machine Type Communication (MTC) and the like. The embodiments of the present disclosure may be performed according to any generation communication protocols either currently known or to be developed in the future. Examples of the communication protocols include, but not limited to, the first generation (1G), the second generation (2G), 2.5G, 2.75G, the third generation (3G), the fourth generation (4G), 4.5G, the fifth generation (5G) communication protocols, 5.5G, 5G-Advanced networks, or the sixth generation (6G) networks.

Example Processes

[0072] Reference is made to FIG. 2A, which illustrates a signaling flow 200A for communication in accordance with some embodiments of the present disclosure. For the purposes of discussion, the signaling flow 200 will be discussed with reference to FIG. 1A and FIG. 1B, for example, by using the first device 110 and the second device 120.

[0073] In the following descriptions, while operations are depicted in a particular order, this should not be understood as requiring that such operations be performed in the particular order shown or in sequential order, or that all illustrated operations be performed, to achieve desirable results. In certain circumstances, multitasking and parallel processing may be advantageous. Likewise, while several specific implementation details are contained in the above discussions, these should not be construed as limitations on the scope of the present disclosure, but rather as descriptions

of features that may be specific to particular embodiments. Certain features that are described in the context of separate embodiments may also be implemented in combination in a single embodiment. Conversely, various features that are described in the context of a single embodiment may also be implemented in multiple embodiments separately or in any suitable sub-combination.

[0074] It is to be understood that the operations at the first device 110 and the second device 120 should be coordinated. In other words, the second device 120 and the first device 110 should have common understanding about configurations, parameters and so on. Such common understanding may be implemented by any suitable interactions between the second device 120 and the first device 110 or both the second device 120 and the first device 110 applying the same rule/policy. In the following, although some operations are described from a perspective of the first device 110, it is to be understood that the corresponding operations should be performed by the second device 120. Similarly, although some operations are described from a perspective of the second device 120, it is to be understood that the corresponding operations should be performed by the first device 110. Merely for brevity, some of the same or similar contents are omitted here.

[0075] In addition, in the following description, some interactions are performed among the first device 110 and the second device 120 (such as, exchanging first and second information and so on). It is to be understood that the interactions may be implemented either in one single signaling/message/configuration or multiple signaling/messages/configurations, including system information, a radio resource control (RRC) signalling, downlink control information (DCI), uplink control information (UCI), media access control (MAC) control element (CE) and so on. The present disclosure is not limited in this regard.

[0076] In some embodiments, the first device 110 may be operated as a terminal device and the second device 120 may be operated as a network device. Alternatively, in some other embodiments, both the first and the second devices are terminal devices.

[0077] According to some embodiments of the present disclosure, the first device 110 may determine a first identity (such as, a UE side maintained model ID if the first device 110 is a UE), where the first identity may be one of the following: a global model identity or a physical model identity. Further, the second device 120 may configure a second

identity for the ML model (such as, a network assigned model ID if the second device 120 is a network), where the second identity may be one of the following: a local model identity or a logical model identity. Alternatively, in some embodiments, the second identity configured for the ML model also may be the ordinal number of the ML model in a configured ML model list.

[0078] In summary, compared with the fist identity, the second identity may have a lower signalling overhead.

[0079] As illustrated in FIG. 2A, optionally, the first device 110 and the second device 120 may perform 220 an identification procedure. With this identification procedure, a mapping between the first identity and the second identity may be established at the first device 110 and the second device 120. In other words, one ML model may correspond to the first identity and the second identity.

[0080] In the following, details about the identification procedure will be described with reference to FIGS. 3A to 3C, which illustrate signaling flows for identification procedures 300A, 300B and 300C in accordance with some embodiments of the present disclosure.

[0081] Reference is now made to FIG. 3A. In operation, the first device 110 may transmit 310 first information to the second device 120, where the first information indicates a set of first identities of a first set of ML models comprising one or more ML models.

[0082] Then the first device 110 may receive 320 second information from the second device 120, where the second information indicates a set of second identities of a second set of ML models.

[0083] Since the first device 110 does not known the second device 120 implementation, and the second device 120 may only select those ML models suitable to configure the second identity. Thus, the set of ML models with the first identity may be larger than the set of ML models with the second identity. That is, in some embodiments, the second set of ML models may be a subset of the first set of ML models.

[0084] As one example, the second device 120 is equipped with 64 Tx beams, the second device 120 may only select ML models trained for 64 Tx beams and configure to first device 110. The other models supporting 16 beams or 128 beams are not useful for this the second device 120.

[0085] By exchanging the first information and the second information, the first device 110 and the second device 120 may obtain mappings between first identities and second identities.

[0086] For example, the set of first identities may be {first identity #1, first identity #2, first identity #3, first identity #4} corresponding to model #1, model #2, model #3 and model #4, respectively, and the set of second identities may be {second identity #1, second identity #2} corresponding to model #1, model #2, respectively. Then the first device 110 and the second device 120 may determine a mapping of {first identity #1, second identity #1} for model #1, and a mapping of {first identity #2, second identity #2} for model #2.

[0087] In some embodiments, the second identity may be determined at least partially based on the first information. In some embodiments, the value of the second identity may be a function of the value of the first identity. In some other embodiments, the first information may contain a second identity(ies) suggested by the first device 110.

[0088] Further, the first identity and the second identity may be used for different LCM procedure set. In some embodiments, the first identity may be used for identifying the ML model during at least of the following procedure: a model transfer procedure, a model delivery procedure, a model update procedure, a model monitoring procedure, a model training procedure, or a data collection procedure. And the second identity may be used for identifying the ML model during at least of the following procedure: a model inference procedure, a model management procedure, a model monitoring procedure, a model training procedure, or a data collection procedure.

[0089] In this way, when the ML model needs to be identified as a physical ML model, the first identify may be used. Else, the second identity may be used for identifying the ML model to reduce the signalling overhead.

[0090] Optionally, in some embodiments, the first device 110 may receive, from the second device 120, a model-related configuration associated with a second identity of the set of second identities, where the model-related configuration is configured for at least one of the following: a model training procedure, a model inference procedure, a model monitoring procedure or a data collection and associated with at least one of the following: measurement, reporting and behaviors of the first device 110.

[0091] In some embodiments, the second identity may be associated with configurations

of measurement, report and UE behavior specified for model training, model inference, model monitoring, data collection and so on. Below table 2 illustrates an example to show the relationship between first and second identity, and the configurations.

Table 2 an example to show the relationship between first and second identities, and the configurations

First identity UE model ID	Second identity	Associated configurations
X1-1, 1	0	Configuration 0: 0^{th} set of resources for measurement, report, and UE behavior
X1-1, 3	1	Configuration 1: 1^{st} set of resources for measurement, report, and UE behavior
X1-1, 5	2	Configuration 2: 2^{nd} set of resources for measurement, report, and UE behavior
X1-1, 7	3	Configuration 3: 3^{rd} set of resources for measurement, report, and UE behavior

[0092] As illustrated in the above table 2, Second identity 0 and First identity "X1-1,1" are confiugued for the ML model 1, further, the Configuration 0 may be associated with the Second identity 0. Further, the Configuration 0 may be configured for at least one of the following: a model training procedure of the ML model 1, a model inference procedure of the ML model 1, a model monitoring procedure or a data collection of the ML model 1 and comprises at least one of : 0^{th} set of resources for measurement, report, and related UE behavior.

[0093] Similarly, in the above table 2, Second identity 1/2/3 and First identity "X1-1,3"/"X1-1,5"/"X1-1,7" are confiugued for the ML model 3/5/7, further, the Configuration 1/2/3 may be associated with the Second identity 1/2/3. Further, the Configuration 1/2/3 may be configured for at least one of the following: a model training procedure of the ML model 3/5/7, a model inference procedure of the ML model 3/5/7, a model monitoring procedure or a data collection of the ML model 3/5/7 and comprises at least one of: $1^{th}/2^{th}3^{th}$ set of resources for measurement, report, and related UE behavior.

[0094] In addition to the example embodiment of FIG. 3A, the identification procedure also may be implemented in other manners.

[0095] Reference is now made to FIG. 3B. In operation, the first device 110 may transmit 330 the first information to the second device 120, where the first information indicates a set of first identities of a first set of ML models comprising the ML model. Then the first device 110 may receive 340 third information from the second device 120, where the third information indicates a set of temporary second identities (which may be a network assigned model ID if the second device 120 is a network device) of a second

set of ML models comprising the ML model, wherein the temporary second identity is configured by the second device 120.

[0096] In this event, the first device 120 needs to determine whether the temporary second identities are acceptable, e.g., the first device 110 needs to check whether the temporary second identity(ies) has been used. Based on the determination, the first device 110 may transmit the feedback information to the second device 120. In some embodiments, the first device 110 may transmit 350 confirmation information which is used for confirming at least one temporary second identity to the second device 120. As the result, the temporality of the temporary second identity may be removed, i.e., the temporary second identity will be used as the second identity.

[0097] Alternatively, the first device 110 may transmit 350 contention information to the second device 120, where the contention information indicates that at least temporary second identity is not applicable. In this event, the second device 120 may optionally transmit 360 contention resolution information to the first device 110, where the contention resolution information indicates at least one newly-configured second identity corresponding to the unacceptable at least temporary second identity.

[0098] Alternatively, the first device 110 may transmit 350 contention information to the second device 120, where the contention information indicates at least one recommended second identity in case that at least one temporary second identity comprised is not applicable. Then, the second device 120 may transmit 360 contention resolution information to the first device 110, where the contention resolution information may indicate whether the at least one recommended second identity is confirmed by the second device 120.

[0099] In this way, by configuring temporary second identity, the first device 110 and the second device 120 also may determine the mapping between the first identity and the second identity.

[0100] A further identification procedure will be discussed with reference to FIG. 3C. In some embodiments, the ML model may be developed at the first device 110. In addition, a third device (a third party, such as, either (or both) AI/ML server for the second device 120 or/and AI/ML server for the first device 110) may deliver an AI/ML model with a third identity to either (or both) of the second device 120 or/and the first device 110.

[0101] As illustrated in FIG. 3C, in operation, prior to communication the first information, the second device 120 may receive 375 a third identity of the ML model configured by the third device, and the first device 110 also may receive 380 a third identity of the ML model configured by the third device.

[0102] In some embodiments, the third identity of a ML model may include one or more of the following information: global model, local model (optional), network ID, UE ID, configuration ID, dataset ID and so on.

[0103] In some embodiments, the third identity may be represented in a hierarchical structure (as illustrated in FIG. 2B) or a concatenation structure (as illustrated in FIG. 2C) to identify a global model.

[0104] In FIG. 2B, the second device 120 may develop three different versions for the ML model. Similarly, the first deice 110 also may develop three different versions for the ML model. As for the mode 0, model 1 and model 2 at the first device 110, they only need to know the mode 0 at the second device 120 , and the other version of the ML model may be transparent for them. That is, the second device 120 may only notify the mode 0 at the second device to the first device 110. In this way, by using this hierarchical structure to identify the ML model, the version maintenance at the related device would be more flexible.

[0105] In FIG. 2C, the identity consists of sequentially connected multiple parts, for example, a global model, a local model, a network (NW) ID, a UE ID, a configuration ID, a dataset ID and so on. In this way, more model-related information may be comprised in the model identity.

[0106] In case of hierarchical structure, the second device 120 and/or the first device 110 may develop different versions of AI/ML models based on the one reference AI/ML model with a global ID. However, the second device 120/the first device 110 versions are transparent to the third party.

[0107] In case of concatenation structure: the second device 120/ the first device 110 versions are not transparent to the third party. Below table 3 illustrates example identities.

Table 3 example identities.

Third identity (Global ID assigned by a third party)	Second identity (Model ID used for activation/de-activation/switching/selection)	First identity (Corresponding to UE model ID)
aaaaaaaa	0	X1-1, 1 (i.e., associated with X1-1 and 1, where X1-1 is an index of a feature group and 1 is an index of the related ML model)
aaaaffff	1	X1-1, 3 (i.e., associated with X1-1 and 1, where X1-1 is an index of a feature group and 3 is an index of the related ML model)
.....	2	X1-1, 5 (i.e., associated with X1-1 and 1, where X1-1 is an index of a feature group and 5 is an index of the related ML model)
ffffffff	3	X1-1, 7 (i.e., associated with X1-1 and 1, where X1-1 is an index of a feature group and 7 is an index of the related ML model)

[0108] In the above table 3, example value of "aaaaaaaa"/"aaaaffff"/ "ffffffff" may be an 8 hexadecimal numbers.

[0109] Still refer to FIG. 3C, in the following, the first device 110 may transmit the first information comprising the third identity to the second device 120. As the third identity is also known to the second device 110, the ML model may be better identified, i.e., the third identity may be used to align the model deployed at second device 120, especially for a two sided model.

[0110] Additionally, after communication of the first information, the first device 110 may receive second information (indicating a set of second identities of a second set of ML models) from the second device 120 as discussed with reference to FIG. 3A, or receive the third information (indicating a set of temporary second identities of a second set of ML models) from the second device 120 as discussed with reference to FIG. 3B 15. Merely for brevity, the same or the similar content are omitted.

[0111] In some embodiments, the first identity and the second identity are transmitted via an RRC signaling, MAC CE, UCI or a capacity report. And the second identity and the second identity may be transmitted via an RRC signaling, MAC CE or DCI.

[0112] In this way, the mapping relationship among different types of AI/ML model identities may be established in model identification procedure. In particular, the second identity may be used to identify an AI/ML model and in subsequent AI/ML operations

between the first device 110 and the second device 120.

[0113] In some embodiments, when transmitting the first identity via a capacity report, the first device 110 may transmit 210 capability-related information of the first device 110 to the second device 120 and via at least one feature or at least one feature group, as illustrated in FIG. 2A. In some embodiments, the capability-related information indicates a set of first identities of a first set of ML models comprising the ML model, wherein the first identity is determined by the first device 110.

[0114] Alternatively, or in addition, except for the set of first identities, the capability-related information also may indicate model-related information, including but not limited to, the number of ML models supported for a specific feature or a feature group, or at least one model-related parameter related to a specific feature or a feature group.

[0115] In some embodiments, a UE capability reporting for AI/ML-enabled feature/feature group is supported, i.e., a UE capability report may include the member of AI/ML models supported for a feature/ feature group and/or details of AI/ML models supported for a feature/feature group. In this way, the first device 110 may indicate supported AI/ML models to the second device 120, and the second device may then perform correct configuration based on the UE capability report. Below table 4 illustrates an example structure of the UE capability report.

Table 4 an example structure of the UE capability report

Features	Index	Feature group	Components	Field name	Candidate values
X1. AI/ML for CSI feedback enhancement	X1-1	CSI compression	1) Supported/max number of input size 1) Nt, number of ports 2) Ns, number of subbands 2) Supported/max number of output size 1) Compression ratio 2) Quantization 3) Payload size for compressed bits 3) Supported AI/ML models 1) Supported/max number 2) Detail information	*AIMLInput_Nt* *AIMLInput_Ns* *AIMLOutput_compressionRatio* *AIMLOutput_quantization* *AIMLOutput_compressedbits* *AIMLModels*	AIMLInput_Nt {4, 8, 16, 32, ...} ...
	...				
X2. AI/ML for time-domain CSI prediction	X2-1	CSI prediction	1) Supported/max number of historical instances 2) Supported/max number of future instances 3) Supported AI/ML models		
	...				
X3. AI/ML for spatial-domain beam prediction	X3-1	DL Tx beam prediction	1) Supported/max size of set A 2) Supported/max size of set B 3) Supported AI/ML models		
	X3-2	Beam pair prediction			
	...				
X4. AI/ML for time-domain beam prediction	X4-1	DL Tx beam prediction	1) Supported/max size of set A 2) Supported/max size of set B 3) Supported/max number of historical instances 4) Supported/max number of future instances 5) Supported AI/ML models		
	X4-2	Beam pair prediction	...		
	...				
X5. AI/ML for direct positioning enhancement	...				
X6. AI/ML assisted positioning enhancement	...				

[0116] The example of above table 4 is given for illustrative purpose only, the values of the items may be changed in the other embodiments. Specifically, in some other embodiments, Feature Index (X1, X2, ...) may be replaced with any integer number. Alternatively, in some other embodiments, Feature Index (X1, X2, ...) may be a type of model identity, functionality identity or a type of the first identity. In some other embodiments, Field names may be other names according to the components of the feature/ feature group. In some other embodiments, the Candidate values may have a different range.

[0117] In the above Table 4, one feature may correspond to more than one Feature group. In some embodiments, each feature may be associated with a sub use case being one of the following: ML model-based channel state information (CSI) feedback, ML model-based time-domain CSI prediction, ML model-based spatial-domain beam prediction, ML model-based time-domain beam prediction, ML model-based direct positioning, or ML model-based assisted positioning. In some embodiments, the feature group may be one of the following: a CSI compression, a CSI prediction, a downlink transmitter beam prediction, or a beam pair prediction, and wherein the capability-related information indicates at least one of the following.

[0118] Further, each feature group may correspond one component in which related parameter associated with this feature group may be configured.

[0119] In some embodiments, the UE capability report may also include: configurations, conditions, additional conditions, UE internal conditions, and so on.

[0120] In some embodiments, details of AI/ML models may be also reported via this feature/feature group report, in addition to the number. As discussed above, a ML model identity or a group of identity may be included, and is associated with different inputs/outputs/other information. Below table 4 illustrates an example using feature/feature group to report AI/ML related UE capability for CSI compression.

Table 5 an example using feature/feature group to report ML related UE capability for CSI compression

Feature index	UE model ID Corresponding to First identity	Associated inputs	Associated outputs	Others associated information

X1-1	1	First set of values of: Nt (number of ports), Ns (number of sub-bands), …	First set of values of compression ratio, quantization and compressed bits, …	SNR > a threshold (i.e., application condition), and so on
X1-1	2	2nd set of values of Nt (number of ports), Ns (number of sub-bands), …	2nd set of values of compression ratio, quantization and compressed bits, …	SNR < a threshold (i.e., application condition), and so on
X1-1	3	3rd set of values of Nt (number of ports), Ns (number of sub-bands), …	3rd set of values of compression ratio, quantization and compressed bits, …	…
…	…	…	…	
X1-1	10	10th set of values of Nt (number of ports), Ns (number of sub-bands), …	10th set of values of compression ratio, quantization and compressed bits, …	

[0121] Through the above processes, the ML model may be well identified by either a first identity or a second identity. However, the LCM of the ML model may be inactive sometimes. In this event, if the configured second identity and related configuration configured for the ML model cannot be released in time, the system resources would be wasted. According to some embodiments of the present disclosure, a timer-based LCM solution is proposed, which will be discussed as below.

[0122] Optionally, the timer-based LCM may be performed only for the model(s) being known to the first device 110. In this event, the first device 110 needs to determine whether the ML model is known to the first device 110. In some embodiments, the first device 110 may determine whether the ML model is known to the first device 110 based on at least of the following: cognition degree on the ML model (such as, model structure or model parameters), availability of model parameters or model structure, or an operation status of the ML mode (such as, a period from last model inference/identification/monitoring).

[0123] In some embodiments, the first device 110 may determine the ML model is known to the first device 110 if:

- at least part of a model structure is known to the first device 110,e.g.,

  the AI/ML model is known if the partial or full model structure is known and/or the partial or full of the model parameters are known;

  for a two-sided model, the AI/ML model is known at the first device 110 if partial or full structure/parameters of UE part of the two sided model are known at the first

device 110;

- at least part of model parameters are known to the first device 110,
- a time length timed from a time point of the last model inference procedure of the ML model is less than or equal to a threshold time, i.e., during a period from the last model inference using this AI/ML model;
- a time length timed from the last model identification procedure of the ML model is less than or equal to a threshold time, i.e., During a period from the last model identification for this AI/ML model;
- a time length timed from the last performance monitoring procedure of the ML model is less than or equal to a threshold time, i.e., during a period from the last performance monitoring of this AI/ML model;
- a time length timed from a completion of the last model training procedure, model fine-tuning procedure or model validation procedure of the ML model is less than or equal to a threshold time, i.e., during a period from the last training/fine-tuning/validation completion of this AI/ML model;
- a time length timed from the last model indication, model measurement, or model reporting related to the ML model is less than or equal to a threshold time, i.e., during a period from the last indication/measurement/reporting related to this AI/ML model;
- a time length timed from the last model activation, switching or selection of the ML model is less than or equal to a threshold time, i.e., during a period from the last model activation/switching-to/selection for this AI/ML model;
- detecting a positive assessment result of the ML model; or
- the availability timer of the ML model is running.

[0124] In some embodiments, the positive performance monitoring results means that satisfy a certain condition, which may be one or many of the following

- Assessment/Monitoring based on the additional conditions associated,
- Assessment/Monitoring based on input/output data distribution,
- Assessment/Monitoring based on inference accuracy/system performance, or

- Assessment/Monitoring based on past knowledge of the performance (e.g., based on other UEs).

[0125] Accordingly, if a condition opposite to the above condition is met, the ML model may be determined to be unknow to the first device 110. Further, during a period from the last model deactivation/switching-from for this ML model, the ML model may be determined to be unknow to the first device 110.

[0126] Accordingly, in some embodiments, in accordance with a determination that the ML model is unknown to the first device 110, the first device 110 may perform at least one of the following:

- requesting another ML model with the second device 120, e.g., requesting a new model, or request to switch to another AI/ML model;
- initiating a model identification procedure for another ML model, or
- applying a longer timer for a model inference procedure of the ML model.

[0127] Still refer to FIG. 2A. In some embodiments, after the identification procedure, the first device 110 may apply 230 one or more identified models. That is, the first device 110 may additionally perform a further selection, for example, based on current application condition, like Signal to Noise Ratio (SNR), mobility, line of sight (LOS)/ non line of sight (NLOS) and so on for example, based on current application condition, like SNR, mobility, LOS/NLOS and so on.

[0128] Alternatively, or in addition, the second device 120 also may transmit 240 a message used for configurating (such as, configuring the second identity and related resource as discussed above) or activating one or more ML models deployed at the first device 110. .Alternatively, or in addition, the second device 120 may switch/select the ML model.

[0129] As a result, one or more ML models may be running at the first device 110. In some embodiments, as illustrated in FIG. 2A, the first device 110 may assess 250 the one or more running ML models, such as, monitoring the model inference/model activation/switching-to/selection/model update/fine-tuning/model retraining, or assessing the performance and so on. Sometimes, the assessment results may be used for maintaining an availability timer for a ML model.

[0130] In the following, details about how to maintain the availability timer for a ML model will be discussed.

[0131] It should be understood that, in the following discussions, although the availability timer is discussed with respect to a ML model, the availability timer may be configured for a group of ML models including one or more ML model. In view of this, all the discussions made to the ML model should also may applicable for the group of ML model, such as, "a positive assessment result of the ML model" may be replaced by "a positive assessment result of at least one ML model in the group of ML model", "a first indication indicating the first device 110 to continue applying the ML model" may be replaced by "a first indication indicating the first device 110 to continue applying at least one ML model in the group of ML model", "a model inference/model activation/switching-to/selection/model update/fine-tuning/model retraining of the ML model" may be replaced by "a model inference/model activation/switching-to/selection/model update/fine-tuning/model retraining of at least one ML model in the group of ML model" and so on.

[0132] As illustrated in FIG. 2A, the first device 110 starts 260 an availability timer for a ML model known to the first device 110.

[0133] Then, upon an expiry of the availability timer, first device 110 transmits 282 a request to the second device 120, where the request indicates the second device 120 to release a model identity configured for the ML model associated with the availability timer.

[0134] In this way, when the LCM of the ML model known to the first device 110 is not active (not in use), the second device 120 may release the model identity configured for the ML model. In this way, a recycling rate pf the available model identity is increased.

[0135] In some embodiments, a time length of the availability timer may be defined as a default value or determined by the first device 110 or the second device 120.

[0136] In some embodiments, a same or different timer(s) may be associated with different ML models/model groups.

[0137] In some embodiments, the first device 110 may start 260 the availability timer upon receiving a first message from the second device 120, where the first message is used for configuring or activating the ML model.

[0138] In some embodiments, upon receiving configuration/activation message from the second device 120, or a server to configure/activate the group of ML models, the first device 110 may start the availability timer. In this event, it implies that the started availability timer is only associated with configured/activated ML models.

[0139] Alternatively, in some embodiments, the first device 110 may start 260 the availability timer upon initiating a model identification procedure for the ML model with the second device 120.

[0140] In some embodiments, the first device 110 may restart 270 the availability timer in response to at least one of the following:

- detecting/determining a positive assessment result of the ML model. In this event, it implies the length of the availability timer should be greater than the time configured for performance monitoring,
- receiving a first indication indicating the first device 110 to continue applying the ML model, such as, receiving an indication from the second device 120/server to apply this AI/ML model for model inference: e.g., model activation/switching-to/selection,
- a start/initiation or a completion of a model update procedure of the ML model,
- a start/initiation or a completion of a model fine-tuning procedure of the ML model, or
- a start/initiation or a completion of a model retraining procedure of the ML model.

[0141] In some embodiments, the first device 110 may determine 280 that the availability timer expiries upon one of the following:

- detecting/determining a negative assessment result of the ML model,
- receiving a second indication indicating the first device 110 not to continue applying the ML model, or after a pre-configured period after receiving the second indication,
- detecting/determining a change of configuration information configured for the ML model,
- detecting/determining a change of an application condition for the ML model at the first device 110, such as, a lacking memory, battery, computation resource,

overheating and other hardware limitations, or

- detecting/determining a predefined event associated with radio resource management, e.g., handover, radio link failure and so on.

[0142] In some embodiments, upon an expiry of the availability timer, the first device 110 may perform at least one the following:

- falling back to a default ML model or an operation mode without ML model,
- reporting the falling back to the second device 120,
- releasing the model identity configured for the ML model,
- releasing a model-related configuration associated with the ML model, the model-related configuration associated with at least one of the following: measurement, reporting and behaviors of the first device 110,
- initiating a model identification procedure to identify another ML model,
- initiating a data collection procedure for the ML model,
- initiating a model update procedure for the ML model,
- initiating a model retraining procedure for the ML model,
- initiating a model deactivation procedure for the ML model, or
- determining that the ML mode is unknown to the first device 110.

[0143] As discussed above, the first device 110 and the second device 120 may release the model identity. In some embodiments, the first device 110 and the second device 120 may release the model identity by releasing a correspondence between a first identity of the ML model and a second identity of the ML model, wherein the second identity is configured by the second device 120 for the ML model and the first identity is determined by the first device 110.

[0144] Alternatively, the first device 110 and the second device 120 may release the model identity by remapping the second identity previously configured for the ML model to another ML model.

[0145] Alternatively, the first device 110 and the second device 120 may release the

model identity by removing the model identity from a list of ML model identity .

[0146] In some embodiments, before and after releasing the model identity, the same second identity may correspond to different AI/ML models. Below table 6 illustrates an example to show that model 5 is released.

table 6 an example to show that model 5 is released

Second identity	First identity Before releasing	First identity After releasing
0	X1-1, 1	X1-1, 1
1	X1-1, 3	X1-1, 3
2	X1-1, 5	X1-1, 7
3	X1-1, 7	…

[0147] In the above table 6, when releasing the model identity, the element "X1-1, 5" is removed from the first identity list. As a result, before releasing the model identity, the second identity 2 is mapped to the model 5, after releasing the model identity, the second identity 2 is mapped to the model 7.

[0148] Optionally, after releasing the model identity, the first device 110/the second device 120 may inform the latest/updated model identity mapping(s) to the second device 120/the first device 110.

[0149] Below table 7 illustrates another example to show that model 5 is released.

Table 7 another example to show that model 5 is released

Second identity	First identity Before releasing	First identity After releasing
0	X1-1, 1	X1-1, 1
1	X1-1, 3	X1-1, 3
2	X1-1, 5	X1-1, 11
3	X1-1, 7	X1-1, 7

[0150] In the above table 7, before releasing the model identity, the second identity 2 is mapped to the model 5, after releasing the model identity, the second identity 2 is mapped to the model 11. That is, the second identity 2 is re-assigned/mapped to a new ML model.

[0151] As one example, after releasing the model identity, a new identification for the model 11 is initiated and the second device 120 re-configure the second identity 2 to the model 11.

[0152] Through the above processes, a second identify may be introduced for identifying the ML model, and thus the signalling overhead is reduced thereby. Further, a model ID would not be permanent any more. In this event, if a ML model is not in use, the associated model ID may be released accordingly. Moreover, due to the timely release, the resources (such as, UE memory/battery) used for maintaining models which are not in use may be saved accordingly.

Example Methods

[0153] FIG. 4 illustrates a flowchart of a communication method 400 implemented at a first device in accordance with some embodiments of the present disclosure. For the purpose of discussion, the method 400 will be described from the perspective of the first device 110 in FIG. 1A.

[0154] At block 410, the first device 110 starts an availability timer for a machine learning (ML) model known to the first device.

[0155] At block 420, upon an expiry of the availability timer, the first device 110 transmits to a second device, a request indicating the second device to release a model identity configured for the ML model associated with the availability timer.

[0156] In some example embodiments, the first device 110 may start the availability timer upon one of the following: receiving a first message from the second device, the first message used for configuring or activating the ML model, or initiating a model identification procedure for the ML model with the second device.

[0157] In some example embodiments, the first device 110 may restart the availability timer upon one of the following: a positive assessment result of the ML model, a first indication indicating the first device to continue applying the ML model, a model update procedure of the ML model, a model fine-tuning procedure of the ML model, or a model retraining procedure of the ML model.

[0158] In some example embodiments, the first device 110 may determine that the availability timer expiries upon one of the following: a negative assessment result of the ML model, a second indication indicating the first device not to continue applying the ML model, a change of configuration information configured for the ML model, a change of

an application condition for the ML model at the first device, or a predefined event associated with radio resource management.

[0159] In some example embodiments, the first device 110 may fall back to a default ML model or an operation mode without ML model, report the falling back to the second device, release the model identity configured for the ML model, release a model-related configuration associated with the ML model, the model-related configuration associated with at least one of the following: measurement, reporting and behaviors of the first device, initiate a model identification procedure to identify another ML model, initiate a data collection procedure for the ML model, initiate a model update procedure for the ML model, initiate a model retraining procedure for the ML model, initiate a model deactivation procedure for the ML model, or determine that the ML mode is unknown to the first device.

[0160] In some example embodiments, the first device 110 may release the model identity by at least one of the following: releasing a correspondence between a first identity of the ML model and a second identity of the ML model, wherein the second identity is configured by the second device for the ML model and the first identity is determined by the first device; remapping the second identity previously configured for the ML model to another ML model; or removing the model identity from a list of ML model identities.

[0161] In some example embodiments, the first device 110 may transmit, to the second device, first information indicating a set of first identities of a first set of ML models comprising the ML model, wherein the first identity is determined by the first device; and receive, from the second device, second information indicating a set of second identities of a second set of ML models comprising the ML model, wherein the second identity is configured by the second device.

[0162] In some example embodiments, the first device 110 may receive, from the second device, a model-related configuration associated with a second identity of the set of second identities, the model-related configuration being configured for at least one of the following: a model training procedure, a model inference procedure, a model monitoring procedure or a data collection and associated with at least one of the following: measurement, reporting and behaviors of the first device.

[0163] In some example embodiments, the second set of ML models is a subset of the

first set of ML models.

[0164] In some example embodiments, the first identity is one of the following: a global model identity or a physical model identity, , or the first identity is used for identifying the ML model during at least of the following procedure: a model transfer procedure, a model delivery procedure, a model update procedure, a model monitoring procedure, a model training procedure, or a data collection procedure.

[0165] In some example embodiments, the second identity is one of the following: a local model identity or a logical model identity, or the second identity is used for identifying the ML model during at least of the following procedure: a model inference procedure, a model management procedure, a model monitoring procedure, a model training procedure, or a data collection procedure.

[0166] In some example embodiments, the first device 110 may transmit, to the second device, first information indicating a set of first identities of a first set of ML models comprising the ML model, wherein the first identity is determined by the first device; receive, from the second device, third information indicating a set of temporary second identities of a second set of ML models comprising the ML model, wherein the temporary second identity is configured by the second device; and transmit, one of the following to the second device: confirmation information used for confirming at least one temporary second identity, contention information indicating at least one the following: at least temporary second identity is not applicable, or at least one recommended second identity in case that at least one temporary second identity comprised is not applicable.

[0167] In some example embodiments, the first device 110 may after transmitting the contention information, receive, from the second device, contention resolution information indicating at least one of the following: whether the at least one recommended second identity is confirmed by the second device, or at least one newly-configured second identity.

[0168] In some example embodiments, the first device 110 may prior to transmitting the first information, receive, from a third device, a third identity of the ML model configured by the third device; and transmit, to the second device, the first information comprising the third identity.

[0169] In some example embodiments, the third identity is represented in a hierarchical

structure or a concatenation structure.

[0170] In some example embodiments, the first device 110 may determine whether the ML model is known to the first device based on at least of the following: cognition degree on the ML model, availability of model parameters or model structure, or an operation status of the ML mode.

[0171] In some example embodiments, the first device 110 may determine the ML model is known to the first device if: at least part of a model structure is known to the first device, at least part of model parameters are known to the first device, a time length timed from a time point of the last model inference procedure of the ML model is less than or equal to a threshold time, a time length timed from the last model identification procedure of the ML model is less than or equal to a threshold time, a time length timed from the last performance monitoring procedure of the ML model is less than or equal to a threshold time, a time length timed from a completion of the last model training procedure, model fine-tuning procedure or model validation procedure of the ML model is less than or equal to a threshold time, a time length timed from the last model indication, model measurement, or model reporting related to the ML model is less than or equal to a threshold time, a time length timed from the last model activation, switching or selection of the ML model is less than or equal to a threshold time, detecting a positive assessment result of the ML model, or the availability timer of the ML model is running.

[0172] In some example embodiments, the first device 110 may in accordance with a determination that the ML model is unknown to the first device, performing at least one of the following: requesting another ML model with the second device, initiating a model identification procedure for another ML model, or applying a longer timer for a model inference procedure of the ML model.

[0173] In some example embodiments, the first device 110 may transmit, to the second device and via at least one feature or at least one feature group, capability-related information of the first device, wherein the capability-related information indicates at least one of the following: a set of first identities of a first set of ML models comprising the ML model, wherein the first identity is determined by the first device, the number of ML models supported for a specific feature or a feature group, or at least one model-related parameter related to a specific feature or a feature group.

[0174] In some example embodiments, a time length of the availability timer is defined

as a default value or determined by the first device or the second device.

[0175] In some example embodiments, the first device is a terminal device and the second device is a network device.

[0176] FIG. 5 illustrates a flowchart of a communication method 500 implemented at a second device in accordance with some embodiments of the present disclosure. For the purpose of discussion, the method 500 will be described from the perspective of the second device 120 in FIG. 1A.

[0177] At block 510, the second device receives, from a first device, a request indicating the second device to release a model identity configured for a ML model known to the first device.

[0178] At block 520, the second device release the model identity based on the request.

[0179] In some example embodiments, releasing the model identity comprising at least one of the following: releasing a correspondence between a first identity of the ML model and a second identity of the ML model, wherein the second identity is configured by the second device for the ML model and the first identity is determined by the first device; remapping the second identity previously configured for the ML model to another ML model; or removing the model identity from a list of ML model identities.

[0180] In some example embodiments, the second device may receive, from the first device, first information indicating a set of first identities of a first set of ML models comprising the ML model, wherein the first identity is determined by the first device; and transmit, to the first device, second information indicating a set of second identities a second set of ML models comprising the ML model, wherein the second identity is configured by the second device for the ML model.

[0181] In some example embodiments, the second device may transmit, to the first device, a model-related configuration configured for at least one of the following: a model training procedure, a model inference procedure, a model monitoring procedure or a data collection and associated with a second identity of the set of second identities, the model-related configuration associated with at least one of the following: measurement, reporting and behavior of the first device.

[0182] In some example embodiments, the second set of ML models is a subset of the first set of ML models.

[0183] In some example embodiments, the first identity is one of the following: a global model identity or a physical model identity, or the first identity is used for identifying the ML model during at least of the following procedure: a model transfer procedure, a model delivery procedure, a model update procedure, a model monitoring procedure, a model training procedure, or a data collection procedure.

[0184] In some example embodiments, the second identity is one of the following: a local model identity or a logical model identity, or the second identity is used for identifying the ML model during at least of the following procedure: a model inference procedure, a model management procedure, a model monitoring procedure, a model training procedure, or a data collection procedure.

[0185] In some example embodiments, the second device may receive, from the first device, first information indicating a set of first identities of a first set of ML models comprising the ML model, wherein the first identity is determined by the first device; transmit, to the first device, third information indicating a set of temporary second identities of a second set of ML models comprising the ML model, wherein the temporary second identity is configured by the second device; and receive, one of the following from the first device: confirmation information used for confirming at least one temporary second identity, contention information indicating at least one the following: at least temporary second identity is not applicable, or at least one recommended second identity in case that at least one temporary second identity comprised is not applicable.

[0186] In some example embodiments, after receiving the contention information the second device may transmit, to the first device contention resolution information indicating at least one of the following: whether the at least one recommended second identity is confirmed by the second device, or at least one newly-configured second identity.

[0187] In some example embodiments, prior to receiving the first information, the second device may receive, from a third device, a fourth identity of the ML model configured by the third device; and receive, from the first device, the first information comprising a third identity of the ML model configured by the third device, wherein the third and the fourth identity are the same or different .

[0188] In some example embodiments, any of the third and fourth identity is represented in a hierarchical structure or a concatenation structure.

[0189] In some example embodiments, the second device may receive, from the first device and via at least one feature or at least one feature group, capability-related information of the first device, wherein the capability-related information indicates at least one of the following: a set of first identities of a first set of ML models comprising the ML model, wherein the first identity is determined by the first device, the number of ML models supported for a specific feature or a feature group, or at least one model-related parameter related to a specific feature or a feature group.

[0190] In some example embodiments, the first device is a terminal device and the second device is a network device.

Example Devices and Apparatuses

[0191] FIG. 6 is a simplified block diagram of a device 600 that is suitable for implementing embodiments of the present disclosure. The device 600 can be considered as a further example implementation of any of the devices as shown in FIG. 1A. Accordingly, the device 600 can be implemented at or as at least a part of the first device 16 or the second device 120.

[0192] As shown, the device 600 includes a processor 610, a memory 620 coupled to the processor 610, a suitable transceiver 640 coupled to the processor 610, and a communication interface coupled to the transceiver 640. The memory 630 stores at least a part of a program 630. The transceiver 640 may be for bidirectional communications or a unidirectional communication based on requirements. The transceiver 640 may include at least one of a transmitter 642 and a receiver 644. The transmitter 642 and the receiver 644 may be functional modules or physical entities. The transceiver 640 has at least one antenna to facilitate communication, though in practice an Access Node mentioned in this application may have several ones. The communication interface may represent any interface that is necessary for communication with other network elements, such as X2/Xn interface for bidirectional communications between eNBs/gNBs, S1/NG interface for communication between a Mobility Management Entity (MME)/Access and Mobility Management Function (AMF)/SGW/UPF and the eNB/gNB, Un interface for communication between the eNB/gNB and a relay node (RN), or Uu interface for communication between the eNB/gNB and a terminal device.

[0193] The program 630 is assumed to include program instructions that, when executed by the associated processor 610, enable the device 600 to operate in accordance with the embodiments of the present disclosure, as discussed herein with reference to FIGS. 1 to 6. The embodiments herein may be implemented by computer software executable by the processor 610 of the device 600, or by hardware, or by a combination of software and hardware. The processor 610 may be configured to implement various embodiments of the present disclosure. Furthermore, a combination of the processor 610 and memory 620 may form processing means 650 adapted to implement various embodiments of the present disclosure.

[0194] The memory 620 may be of any type suitable to the local technical network and may be implemented using any suitable data storage technology, such as a non-transitory computer readable storage medium, semiconductor based memory devices, magnetic memory devices and systems, optical memory devices and systems, fixed memory and removable memory, as non-limiting examples. While only one memory 620 is shown in the device 600, there may be several physically distinct memory modules in the device 600. The processor 610 may be of any type suitable to the local technical network, and may include one or more of general purpose computers, special purpose computers, microprocessors, digital signal processors (DSPs) and processors based on multicore processor architecture, as non-limiting examples. The device 600 may have multiple processors, such as an application specific integrated circuit chip that is slaved in time to a clock which synchronizes the main processor.

[0195] According to embodiments of the present disclosure, a first device comprising a circuitry is provided. The circuitry is configured to: start an availability timer for a machine learning (ML) model known to the first device; and upon an expiry of the availability timer, transmit, to a second device, a request indicating the second device to release a model identity configured for the ML model associated with the availability timer. According to embodiments of the present disclosure, the circuitry may be configured to perform any method implemented by the first device as discussed above.

[0196] According to embodiments of the present disclosure, a second device comprising a circuitry is provided. The circuitry is configured to: receive, from a first device, a request indicating the second device to release a model identity configured for a ML model known to the first device; and release the model identity based on the request. According to embodiments of the present disclosure, the circuitry may be configured to perform any

method implemented by the second device as discussed above.

[0197] The term "circuitry" used herein may refer to hardware circuits and/or combinations of hardware circuits and software. For example, the circuitry may be a combination of analog and/or digital hardware circuits with software/firmware. As a further example, the circuitry may be any portions of hardware processors with software including digital signal processor(s), software, and memory(ies) that work together to cause an apparatus, such as a terminal device or a network device, to perform various functions. In a still further example, the circuitry may be hardware circuits and or processors, such as a microprocessor or a portion of a microprocessor, that requires software/firmware for operation, but the software may not be present when it is not needed for operation. As used herein, the term circuitry also covers an implementation of merely a hardware circuit or processor(s) or a portion of a hardware circuit or processor(s) and its (or their) accompanying software and/or firmware.

[0198] According to embodiments of the present disclosure, a first apparatus is provided. The first apparatus comprises means for starting an availability timer for a machine learning (ML) model known to the first device; and means for upon an expiry of the availability timer, transmitting, to a second device, a request indicating the second device to release a model identity configured for the ML model associated with the availability timer. In some embodiments, the first apparatus may comprise means for performing the respective operations of the method 400. In some example embodiments, the first apparatus may further comprise means for performing other operations in some example embodiments of the method 400. The means may be implemented in any suitable form. For example, the means may be implemented in a circuitry or software module.

[0199] According to embodiments of the present disclosure, a second apparatus is provided. The second apparatus comprises means for receiving, from a first device, a request indicating the second device to release a model identity configured for a ML model known to the first device; and means for releasing the model identity based on the request. In some embodiments, the second apparatus may comprise means for performing the respective operations of the method 500. In some example embodiments, the second apparatus may further comprise means for performing other operations in some example embodiments of the method 500. The means may be implemented in any suitable form. For example, the means may be implemented in a circuitry or software module.

[0200] In summary, embodiments of the present disclosure provide the following aspects.

[0201] In an aspect, it is proposed a first device comprising: a processor configured to cause the first device to: start an availability timer for a machine learning (ML) model known to the first device; and upon an expiry of the availability timer, transmit, to a second device, a request indicating the second device to release a model identity configured for the ML model associated with the availability timer.

[0202] In some embodiments, the processor is further configured to cause the first device to: start the availability timer upon one of the following: receiving a first message from the second device, the first message used for configuring or activating the ML model, or initiating a model identification procedure for the ML model with the second device.

[0203] In some embodiments, the processor is further configured to cause the first device to: restart the availability timer upon one of the following: a positive assessment result of the ML model, a first indication indicating the first device to continue applying the ML model, a model update procedure of the ML model, a model fine-tuning procedure of the ML model, or a model retraining procedure of the ML model.

[0204] In some embodiments, the processor is further configured to cause the first device to: determine that the availability timer expiries upon one of the following: a negative assessment result of the ML model, a second indication indicating the first device not to continue applying the ML model, a change of configuration information configured for the ML model, a change of an application condition for the ML model at the first device, or a predefined event associated with radio resource management.

[0205] In some embodiments, the processor is further configured to cause the first device to: fall back to a default ML model or an operation mode without ML model, report the falling back to the second device, release the model identity configured for the ML model, release a model-related configuration associated with the ML model, the model-related configuration associated with at least one of the following: measurement, reporting and behaviors of the first device, initiate a model identification procedure to identify another ML model, initiate a data collection procedure for the ML model, initiate a model update procedure for the ML model, initiate a model retraining procedure for the ML model, initiate a model deactivation procedure for the ML model, or determine that the ML mode is unknown to the first device.

[0206] In some embodiments, the processor is further configured to cause the first device to: release the model identity by at least one of the following: releasing a correspondence between a first identity of the ML model and a second identity of the ML model, wherein the second identity is configured by the second device for the ML model and the first identity is determined by the first device; remapping the second identity previously configured for the ML model to another ML model; or removing the model identity from a list of ML model identities.

[0207] In some embodiments, the processor is further configured to cause the first device to: transmit, to the second device, first information indicating a set of first identities of a first set of ML models comprising the ML model, wherein the first identity is determined by the first device; and receive, from the second device, second information indicating a set of second identities of a second set of ML models comprising the ML model, wherein the second identity is configured by the second device.

[0208] In some embodiments, the processor is further configured to cause the first device to: receive, from the second device, a model-related configuration associated with a second identity of the set of second identities, the model-related configuration being configured for at least one of the following: a model training procedure, a model inference procedure, a model monitoring procedure or a data collection and associated with at least one of the following: measurement, reporting and behaviors of the first device.

[0209] In some embodiments, the second set of ML models is a subset of the first set of ML models.

[0210] In some embodiments, the first identity is one of the following: a global model identity or a physical model identity, , or the first identity is used for identifying the ML model during at least of the following procedure: a model transfer procedure, a model delivery procedure, a model update procedure, a model monitoring procedure, a model training procedure, or a data collection procedure.

[0211] In some embodiments, the second identity is one of the following: a local model identity or a logical model identity, or the second identity is used for identifying the ML model during at least of the following procedure: a model inference procedure, a model management procedure, a model monitoring procedure, a model training procedure, or a data collection procedure.

[0212] In some embodiments, the processor is further configured to cause the first device to: transmit, to the second device, first information indicating a set of first identities of a first set of ML models comprising the ML model, wherein the first identity is determined by the first device; receive, from the second device, third information indicating a set of temporary second identities of a second set of ML models comprising the ML model, wherein the temporary second identity is configured by the second device; and transmit, one of the following to the second device: confirmation information used for confirming at least one temporary second identity, contention information indicating at least one the following: at least temporary second identity is not applicable, or at least one recommended second identity in case that at least one temporary second identity comprised is not applicable.

[0213] In some embodiments, the processor is further configured to cause the first device to: after transmitting the contention information, receive, from the second device, contention resolution information indicating at least one of the following: whether the at least one recommended second identity is confirmed by the second device, or at least one newly-configured second identity.

[0214] In some embodiments, the processor is further configured to cause the first device to: prior to transmitting the first information, receive, from a third device, a third identity of the ML model configured by the third device; and transmit, to the second device, the first information comprising the third identity.

[0215] In some embodiments, the third identity is represented in a hierarchical structure or a concatenation structure.

[0216] In some embodiments, the processor is further configured to cause the first device to: determine whether the ML model is known to the first device based on at least of the following: cognition degree on the ML model, availability of model parameters or model structure, or an operation status of the ML mode.

[0217] In some embodiments, the processor is further configured to cause the first device to: determine the ML model is known to the first device if: at least part of a model structure is known to the first device, at least part of model parameters are known to the first device, a time length timed from a time point of the last model inference procedure of the ML model is less than or equal to a threshold time, a time length timed from the last model identification procedure of the ML model is less than or equal to a threshold

time, a time length timed from the last performance monitoring procedure of the ML model is less than or equal to a threshold time, a time length timed from a completion of the last model training procedure, model fine-tuning procedure or model validation procedure of the ML model is less than or equal to a threshold time, a time length timed from the last model indication, model measurement, or model reporting related to the ML model is less than or equal to a threshold time, a time length timed from the last model activation, switching or selection of the ML model is less than or equal to a threshold time, detecting a positive assessment result of the ML model, or the availability timer of the ML model is running.

[0218] In some embodiments, the processor is further configured to cause the first device to: in accordance with a determination that the ML model is unknown to the first device, performing at least one of the following: requesting another ML model with the second device, initiating a model identification procedure for another ML model, or applying a longer timer for a model inference procedure of the ML model.

[0219] In some embodiments, the processor is further configured to cause the first device to: transmit, to the second device and via at least one feature or at least one feature group, capability-related information of the first device, wherein the capability-related information indicates at least one of the following: a set of first identities of a first set of ML models comprising the ML model, wherein the first identity is determined by the first device, the number of ML models supported for a specific feature or a feature group, or at least one model-related parameter related to a specific feature or a feature group.

[0220] In some embodiments, a time length of the availability timer is defined as a default value or determined by the first device or the second device.

[0221] In some embodiments, the first device is a terminal device and the second device is a network device.

[0222] In an aspect, it is proposed a second device comprising: a processor configured to cause the second device to: receive, from a first device, a request indicating the second device to release a model identity configured for a ML model known to the first device; and release the model identity based on the request.

[0223] In some embodiments, releasing the model identity comprising at least one of the following: releasing a correspondence between a first identity of the ML model and a

second identity of the ML model, wherein the second identity is configured by the second device for the ML model and the first identity is determined by the first device; remapping the second identity previously configured for the ML model to another ML model; or removing the model identity from a list of ML model identities.

[0224] In some embodiments, the processor is further configured to cause the second device to: receive, from the first device, first information indicating a set of first identities of a first set of ML models comprising the ML model, wherein the first identity is determined by the first device; and transmit, to the first device, second information indicating a set of second identities a second set of ML models comprising the ML model, wherein the second identity is configured by the second device for the ML model.

[0225] In some embodiments, the processor is further configured to cause the second device to: transmit, to the first device, a model-related configuration configured for at least one of the following: a model training procedure, a model inference procedure, a model monitoring procedure or a data collection and associated with a second identity of the set of second identities, the model-related configuration associated with at least one of the following: measurement, reporting and behavior of the first device.

[0226] In some embodiments, the second set of ML models is a subset of the first set of ML models.

[0227] In some embodiments, the first identity is one of the following: a global model identity or a physical model identity, or the first identity is used for identifying the ML model during at least of the following procedure: a model transfer procedure, a model delivery procedure, a model update procedure, a model monitoring procedure, a model training procedure, or a data collection procedure.

[0228] In some embodiments, the second identity is one of the following: a local model identity or a logical model identity, or the second identity is used for identifying the ML model during at least of the following procedure: a model inference procedure, a model management procedure, a model monitoring procedure, a model training procedure, or a data collection procedure.

[0229] In some embodiments, the processor is further configured to cause the second device to: receive, from the first device, first information indicating a set of first identities of a first set of ML models comprising the ML model, wherein the first identity is

determined by the first device; transmit, to the first device, third information indicating a set of temporary second identities of a second set of ML models comprising the ML model, wherein the temporary second identity is configured by the second device; and receive, one of the following from the first device: confirmation information used for confirming at least one temporary second identity, contention information indicating at least one the following: at least temporary second identity is not applicable, or at least one recommended second identity in case that at least one temporary second identity comprised is not applicable.

[0230] In some embodiments, the processor is further configured to cause the second device to: after receiving the contention information, transmit, to the first device contention resolution information indicating at least one of the following: whether the at least one recommended second identity is confirmed by the second device, or at least one newly-configured second identity.

[0231] In some embodiments, the processor is further configured to cause the first device to: prior to receiving the first information, receive, from a third device, a fourth identity of the ML model configured by the third device; and receive, from the first device, the first information comprising a third identity of the ML model configured by the third device, wherein the third and the fourth identity are the same or different .

[0232] In some embodiments, any of the third and fourth identity is represented in a hierarchical structure or a concatenation structure.

[0233] In some embodiments, the processor is further configured to cause the second device to: receive, from the first device and via at least one feature or at least one feature group, capability-related information of the first device, wherein the capability-related information indicates at least one of the following: a set of first identities of a first set of ML models comprising the ML model, wherein the first identity is determined by the first device, the number of ML models supported for a specific feature or a feature group, or at least one model-related parameter related to a specific feature or a feature group.

[0234] In some embodiments, the first device is a terminal device and the second device is a network device.

[0235] In an aspect, a first device comprises: at least one processor; and at least one memory coupled to the at least one processor and storing instructions thereon, the

instructions, when executed by the at least one processor, causing the device to perform the method implemented by the first device discussed above.

[0236] In an aspect, a second device comprises: at least one processor; and at least one memory coupled to the at least one processor and storing instructions thereon, the instructions, when executed by the at least one processor, causing the device to perform the method implemented by the second device discussed above.

[0237] In an aspect, a computer readable medium having instructions stored thereon, the instructions, when executed on at least one processor, causing the at least one processor to perform the method implemented by the first device discussed above.

[0238] In an aspect, a computer readable medium having instructions stored thereon, the instructions, when executed on at least one processor, causing the at least one processor to perform the method implemented by the second device discussed above.

[0239] In an aspect, a computer program comprising instructions, the instructions, when executed on at least one processor, causing the at least one processor to perform the method implemented by the first device discussed above.

[0240] In an aspect, a computer program comprising instructions, the instructions, when executed on at least one processor, causing the at least one processor to perform the method implemented by the second device discussed above.

[0241] Generally, various embodiments of the present disclosure may be implemented in hardware or special purpose circuits, software, logic or any combination thereof. Some aspects may be implemented in hardware, while other aspects may be implemented in firmware or software which may be executed by a controller, microprocessor or other computing device. While various aspects of embodiments of the present disclosure are illustrated and described as block diagrams, flowcharts, or using some other pictorial representation, it will be appreciated that the blocks, apparatus, systems, techniques or methods described herein may be implemented in, as non-limiting examples, hardware, software, firmware, special purpose circuits or logic, general purpose hardware or controller or other computing devices, or some combination thereof.

[0242] The present disclosure also provides at least one computer program product tangibly stored on a non-transitory computer readable storage medium. The computer program product includes computer-executable instructions, such as those included in

program modules, being executed in a device on a target real or virtual processor, to carry out the process or method as described above with reference to FIGS. 1 to 6. Generally, program modules include routines, programs, libraries, objects, classes, components, data structures, or the like that perform particular tasks or implement particular abstract data types. The functionality of the program modules may be combined or split between program modules as desired in various embodiments. Machine-executable instructions for program modules may be executed within a local or distributed device. In a distributed device, program modules may be located in both local and remote storage media.

[0243] Program code for carrying out methods of the present disclosure may be written in any combination of one or more programming languages. These program codes may be provided to a processor or controller of a general purpose computer, special purpose computer, or other programmable data processing apparatus, such that the program codes, when executed by the processor or controller, cause the functions/operations specified in the flowcharts and/or block diagrams to be implemented. The program code may execute entirely on a machine, partly on the machine, as a stand-alone software package, partly on the machine and partly on a remote machine or entirely on the remote machine or server.

[0244] The above program code may be embodied on a machine readable medium, which may be any tangible medium that may contain, or store a program for use by or in connection with an instruction execution system, apparatus, or device. The machine readable medium may be a machine readable signal medium or a machine readable storage medium. A machine readable medium may include but not limited to an electronic, magnetic, optical, electromagnetic, infrared, or semiconductor system, apparatus, or device, or any suitable combination of the foregoing. More specific examples of the machine readable storage medium would include an electrical connection having one or more wires, a portable computer diskette, a hard disk, a random access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM or Flash memory), an optical fiber, a portable compact disc read-only memory (CD-ROM), an optical storage device, a magnetic storage device, or any suitable combination of the foregoing.

[0245] Further, while operations are depicted in a particular order, this should not be understood as requiring that such operations be performed in the particular order shown or in sequential order, or that all illustrated operations be performed, to achieve desirable results. In certain circumstances, multitasking and parallel processing may be

advantageous. Likewise, while several specific implementation details are contained in the above discussions, these should not be construed as limitations on the scope of the present disclosure, but rather as descriptions of features that may be specific to particular embodiments. Certain features that are described in the context of separate embodiments may also be implemented in combination in a single embodiment. Conversely, various features that are described in the context of a single embodiment may also be implemented in multiple embodiments separately or in any suitable sub-combination.

[0246] Although the present disclosure has been described in language specific to structural features and/or methodological acts, it is to be understood that the present disclosure defined in the appended claims is not necessarily limited to the specific features or acts described above. Rather, the specific features and acts described above are disclosed as example forms of implementing the claims.

CLAIMS

1. A first device comprising:

a processor configured to cause the first device to:

start an availability timer for a machine learning (ML) model known to the first device; and

upon an expiry of the availability timer, transmit, to a second device, a request indicating the second device to release a model identity configured for the ML model associated with the availability timer.

2. The first device of claim 1, wherein the processor is further configured to cause the first device to:

start the availability timer upon one of the following:

receiving a first message from the second device, the first message used for configuring or activating the ML model, or

initiating a model identification procedure for the ML model with the second device.

3. The first device of claim 1, wherein the processor is further configured to cause the first device to:

restart the availability timer upon one of the following:

a positive assessment result of the ML model,

a first indication indicating the first device to continue applying the ML model,

a model update procedure of the ML model,

a model fine-tuning procedure of the ML model, or

a model retraining procedure of the ML model.

4. The first device of claim 1, wherein the processor is further configured to cause the first device to:

determine that the availability timer expiries upon one of the following:

a negative assessment result of the ML model,

a second indication indicating the first device not to continue applying the ML model,

a change of configuration information configured for the ML model,

a change of an application condition for the ML model at the first device, or

a predefined event associated with radio resource management.

5. The first device of claim 1, wherein the processor is further configured to cause the first device to:

fall back to a default ML model or an operation mode without ML model,

report the falling back to the second device,

release the model identity configured for the ML model,

release a model-related configuration associated with the ML model, the model-related configuration associated with at least one of the following: measurement, reporting and behaviors of the first device,

initiate a model identification procedure to identify another ML model,

initiate a data collection procedure for the ML model,

initiate a model update procedure for the ML model,

initiate a model retraining procedure for the ML model,

initiate a model deactivation procedure for the ML model, or

determine that the ML mode is unknown to the first device.

6. The first device of claim 5, wherein the processor is further configured to cause the first device to:

release the model identity by at least one of the following:

releasing a correspondence between a first identity of the ML model and a second identity of the ML model, wherein the second identity is configured by the second device for the ML model and the first identity is determined by the first device;

remapping the second identity previously configured for the ML model to another ML model; or

removing the model identity from a list of ML model identities.

7. The first device of claim 1, wherein the processor is further configured to cause the first device to:

transmit, to the second device, first information indicating a set of first identities of a first set of ML models comprising the ML model, wherein the first identity is determined by the first device; and

receive, from the second device, second information indicating a set of second identities

of a second set of ML models comprising the ML model, wherein the second identity is configured by the second device.

8. The first device of claim 7, wherein the processor is further configured to cause the first device to:

receive, from the second device, a model-related configuration associated with a second identity of the set of second identities, the model-related configuration being configured for at least one of the following: a model training procedure, a model inference procedure, a model monitoring procedure or a data collection and associated with at least one of the following: measurement, reporting and behaviors of the first device.

9. The first device of claim 1, wherein the processor is further configured to cause the first device to:

transmit, to the second device, first information indicating a set of first identities of a first set of ML models comprising the ML model, wherein the first identity is determined by the first device;

receive, from the second device, third information indicating a set of temporary second identities of a second set of ML models comprising the ML model, wherein the temporary second identity is configured by the second device; and

transmit, one of the following to the second device:

confirmation information used for confirming at least one temporary second identity,

contention information indicating at least one the following:

at least temporary second identity is not applicable, or

at least one recommended second identity in case that at least one temporary second identity comprised is not applicable.

10. The first device of claim 9, wherein the processor is further configured to cause the first device to:

after transmitting the contention information, receive, from the second device, contention resolution information indicating at least one of the following:

whether the at least one recommended second identity is confirmed by the second device, or

at least one newly-configured second identity.

11. The first device of claim 7 or 9, wherein the processor is further configured to cause the first device to:

prior to transmitting the first information, receive, from a third device, a third identity of the ML model configured by the third device; and

transmit, to the second device, the first information comprising the third identity.

12. The first device of claim 11, wherein the third identity is represented in a hierarchical structure or a concatenation structure.

13. The first device of claim 1, wherein the processor is further configured to cause the first device to:

determine whether the ML model is known to the first device based on at least of the following:

cognition degree on the ML model,

availability of model parameters or model structure, or

an operation status of the ML mode.

14. The first device of claim 1, wherein the processor is further configured to cause the first device to:

determine the ML model is known to the first device if:

at least part of a model structure is known to the first device,

at least part of model parameters are known to the first device,

a time length timed from a time point of the last model inference procedure of the ML model is less than or equal to a threshold time,

a time length timed from the last model identification procedure of the ML model is less than or equal to a threshold time,

a time length timed from the last performance monitoring procedure of the ML model is less than or equal to a threshold time,

a time length timed from a completion of the last model training procedure, model fine-tuning procedure or model validation procedure of the ML model is less than or equal to a threshold time,

a time length timed from the last model indication, model measurement, or model reporting related to the ML model is less than or equal to a threshold time,

a time length timed from the last model activation, switching or selection of the ML model is less than or equal to a threshold time,

detecting a positive assessment result of the ML model, or

the availability timer of the ML model is running.

15. The first device of claim 1, wherein the processor is further configured to cause the first device to:

in accordance with a determination that the ML model is unknown to the first device, perform at least one of the following:

requesting another ML model with the second device,

initiating a model identification procedure for another ML model, or

applying a longer timer for a model inference procedure of the ML model.

16. The first device of claim 1, wherein the processor is further configured to cause the first device to:

transmit, to the second device and via at least one feature or at least one feature group, capability-related information of the first device, wherein the capability-related information indicates at least one of the following:

a set of first identities of a first set of ML models comprising the ML model, wherein the first identity is determined by the first device,

the number of ML models supported for a specific feature or a feature group, or

at least one model-related parameter related to a specific feature or a feature group.

17. A second device comprising:

a processor configured to cause the second device to:

receive, from a first device, a request indicating the second device to release a model identity configured for a ML model known to the first device; and

release the model identity based on the request.

18. The second device of claim 17, wherein releasing the model identity comprising at least one of the following:

releasing a correspondence between a first identity of the ML model and a second identity of the ML model, wherein the second identity is configured by the second device for

the ML model and the first identity is determined by the first device;

remapping the second identity previously configured for the ML model to another ML model; or

removing the model identity from a list of ML model identities.

19. The second device of claim 17, wherein the processor is further configured to cause the second device to:

receive, from the first device, first information indicating a set of first identities of a first set of ML models comprising the ML model, wherein the first identity is determined by the first device; and

transmit, to the first device, second information indicating a set of second identities a second set of ML models comprising the ML model, wherein the second identity is configured by the second device for the ML model.

20. A communication method implemented at a first device, comprising:

starting an availability timer for a machine learning (ML) model known to the first device; and

upon an expiry of the availability timer, transmitting, to a second device, a request indicating the second device to release a model identity configured for the ML model associated with the availability timer.

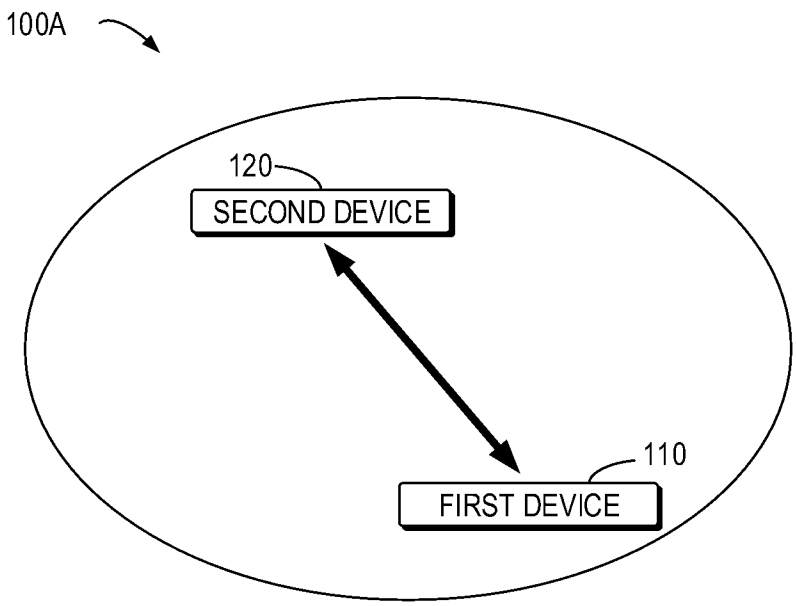
100A
120
SECOND DEVICE
110
FIRST DEVICE

FIG. 1A

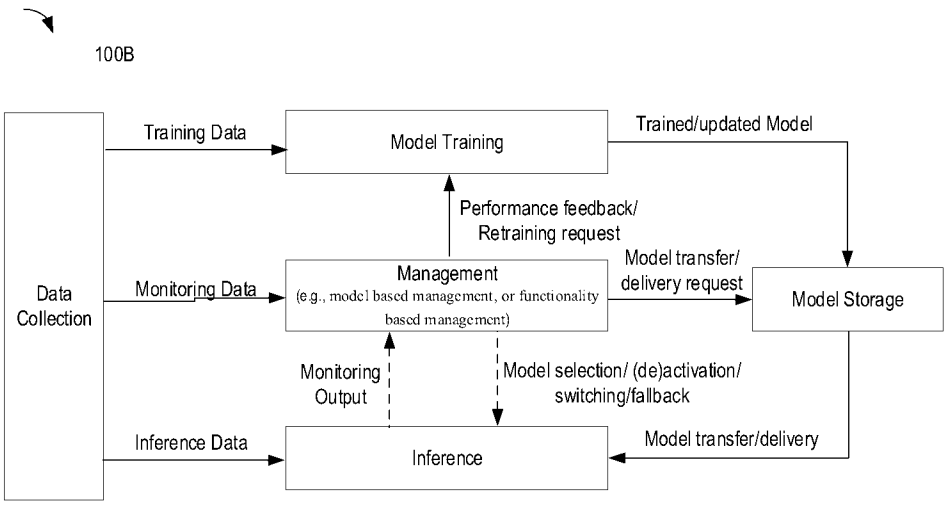
100B
Data Collection
Training Data
Model Training
Trained/updated Model
Performance feedback/ Retraining request
Monitoring Data
Management (e.g., model based management, or functionality based management)
Model transfer/ delivery request
Model Storage
Monitoring Output
Model selection/ (de)activation/ switching/fallback
Inference Data
Inference
Model transfer/delivery

FIG. 1B

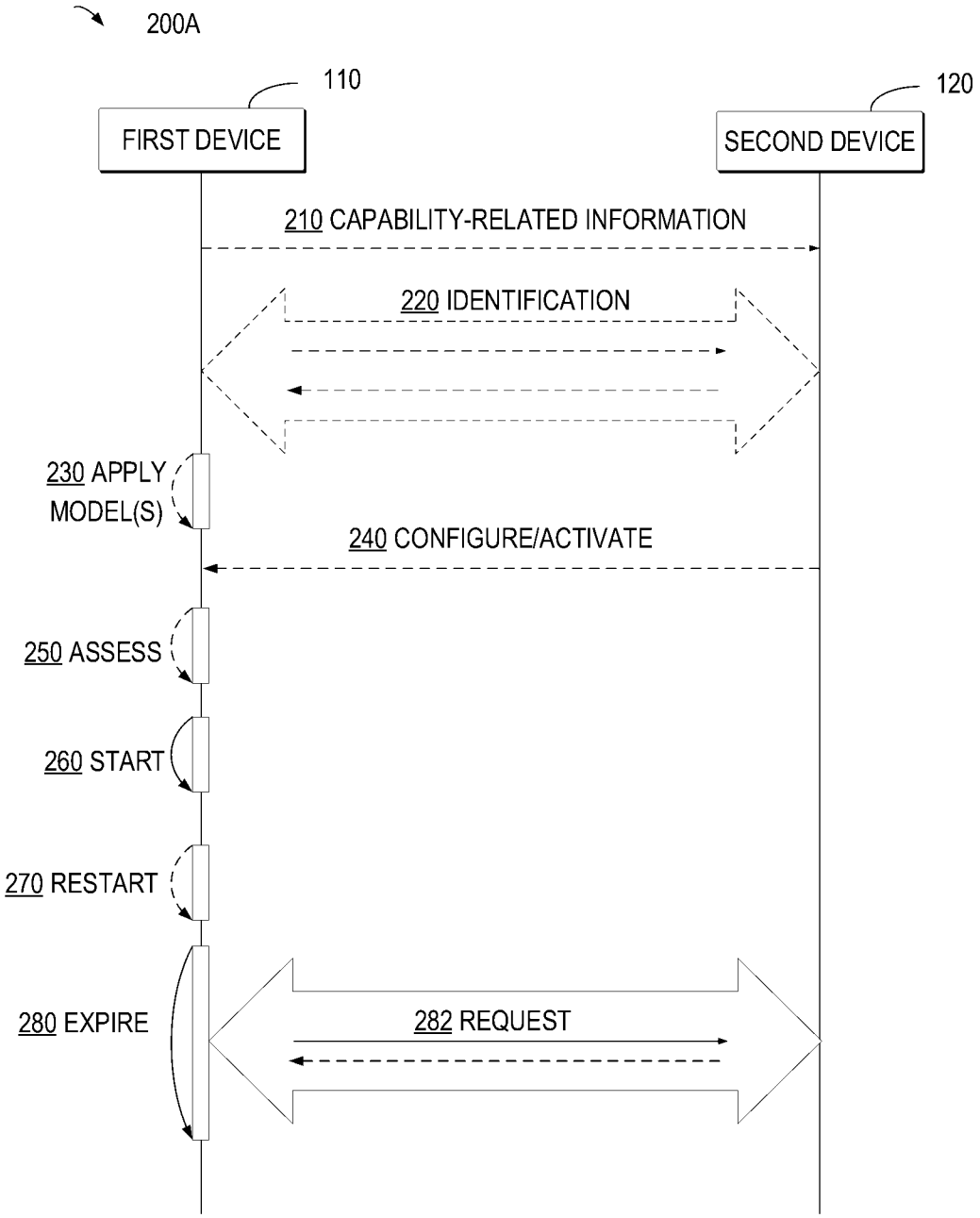
200A
110
120
FIRST DEVICE
SECOND DEVICE
210 CAPABILITY-RELATED INFORMATION
220 IDENTIFICATION
230 APPLY MODEL(S)
240 CONFIGURE/ACTIVATE
250 ASSESS
260 START
270 RESTART
280 EXPIRE
282 REQUEST

FIG. 2A

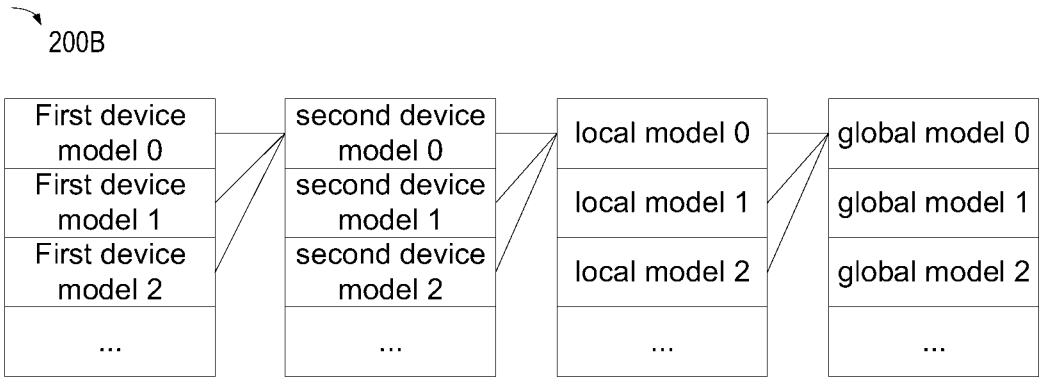
200B
First device model 0
First device model 1
First device model 2
...
second device model 0
second device model 1
second device model 2
...
local model 0
local model 1
local model 2
...
global model 0
global model 1
global model 2
...

FIG. 2B

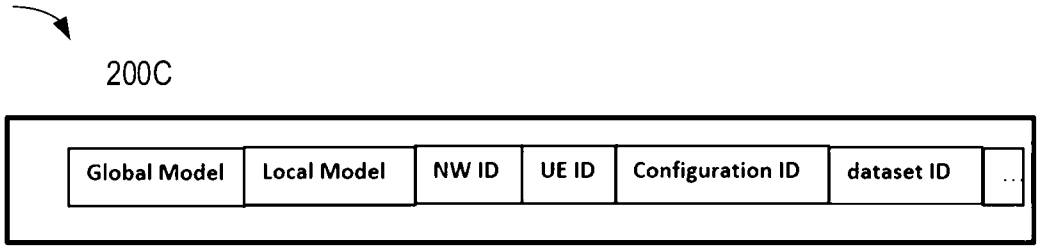
200C
Global Model
Local Model
NW ID
UE ID
Configuration ID
dataset ID
...

FIG. 2C

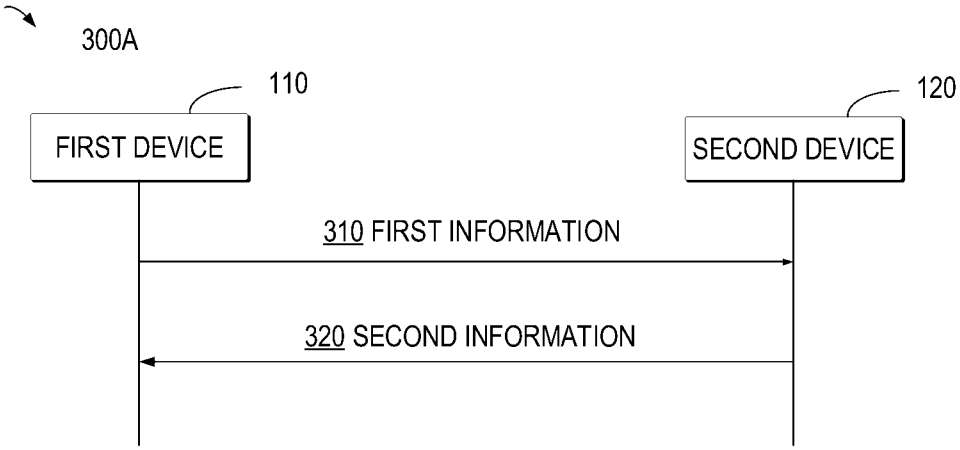
300A
110
120
FIRST DEVICE
SECOND DEVICE
310 FIRST INFORMATION
320 SECOND INFORMATION

FIG. 3A

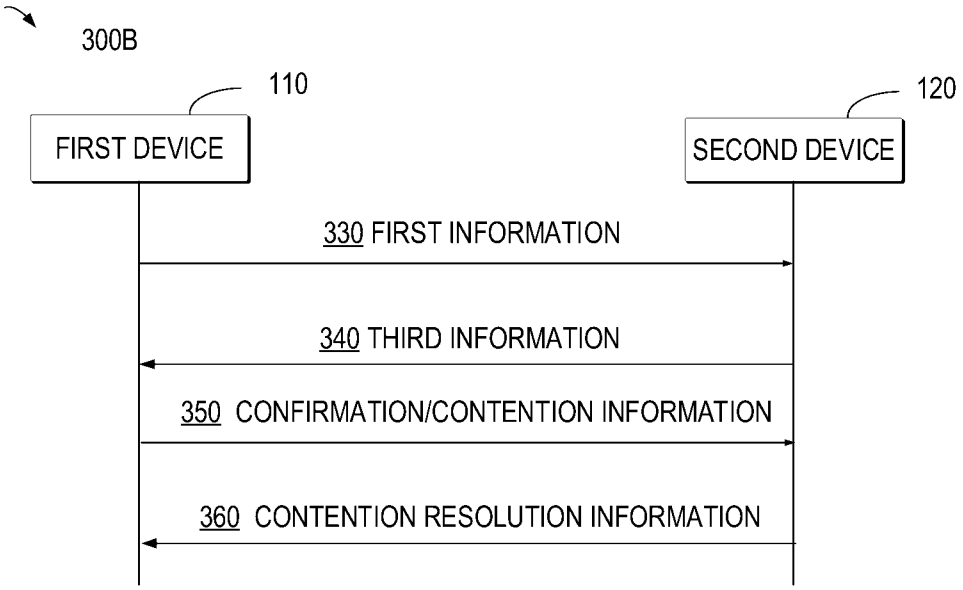
300B
110
120
FIRST DEVICE
SECOND DEVICE
330 FIRST INFORMATION
340 THIRD INFORMATION
350 CONFIRMATION/CONTENTION INFORMATION
360 CONTENTION RESOLUTION INFORMATION

FIG. 3B

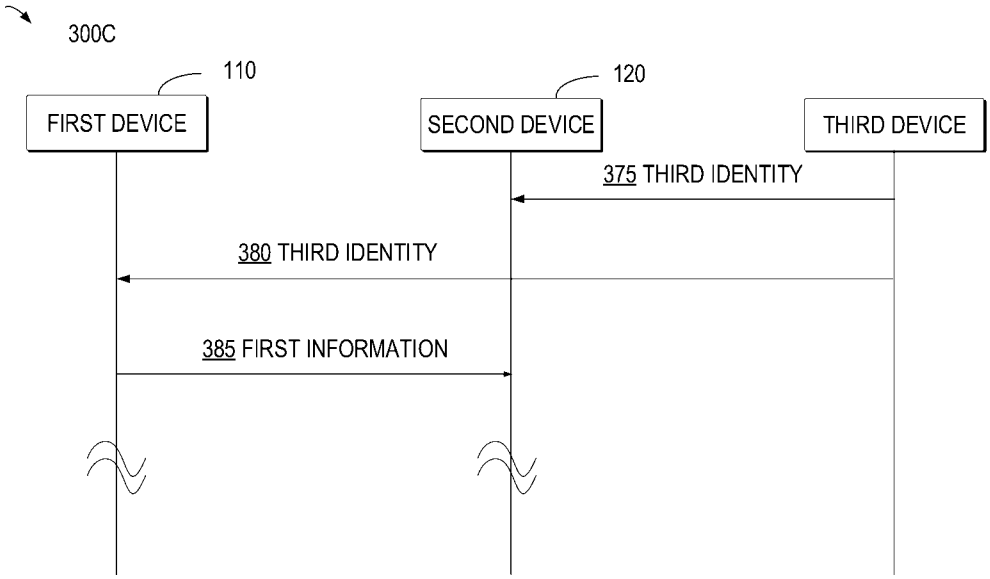
300C
110
120
FIRST DEVICE
SECOND DEVICE
THIRD DEVICE
375 THIRD IDENTITY
380 THIRD IDENTITY
385 FIRST INFORMATION

FIG. 3C

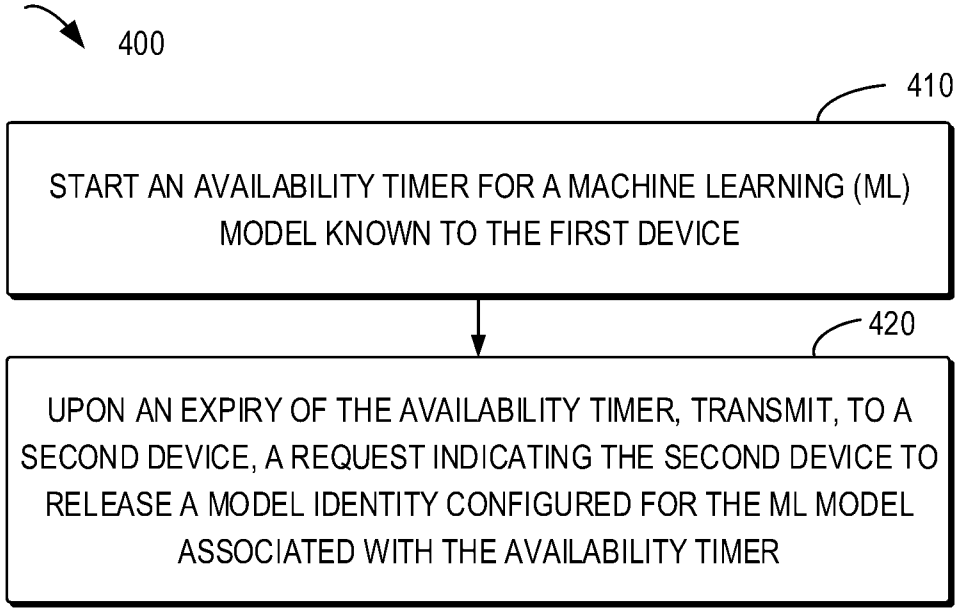
400
410
START AN AVAILABILITY TIMER FOR A MACHINE LEARNING (ML) MODEL KNOWN TO THE FIRST DEVICE
420
UPON AN EXPIRY OF THE AVAILABILITY TIMER, TRANSMIT, TO A SECOND DEVICE, A REQUEST INDICATING THE SECOND DEVICE TO RELEASE A MODEL IDENTITY CONFIGURED FOR THE ML MODEL ASSOCIATED WITH THE AVAILABILITY TIMER

FIG. 4

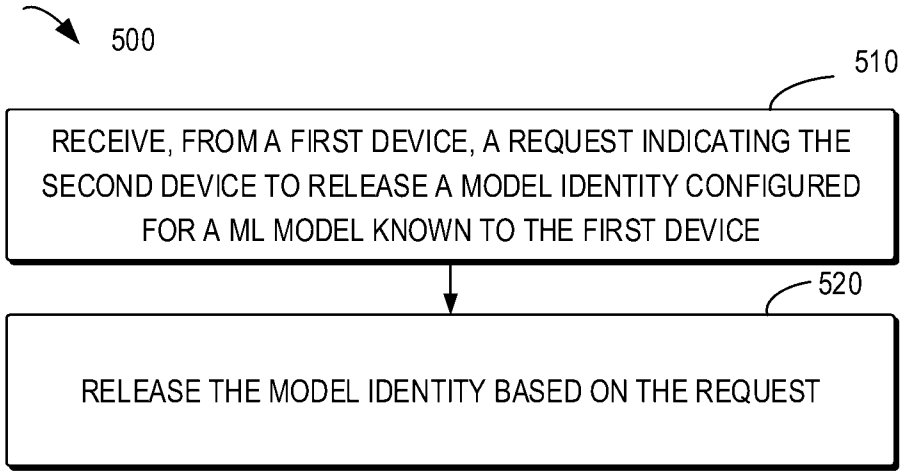
500
510
RECEIVE, FROM A FIRST DEVICE, A REQUEST INDICATING THE SECOND DEVICE TO RELEASE A MODEL IDENTITY CONFIGURED FOR A ML MODEL KNOWN TO THE FIRST DEVICE
520
RELEASE THE MODEL IDENTITY BASED ON THE REQUEST

FIG. 5

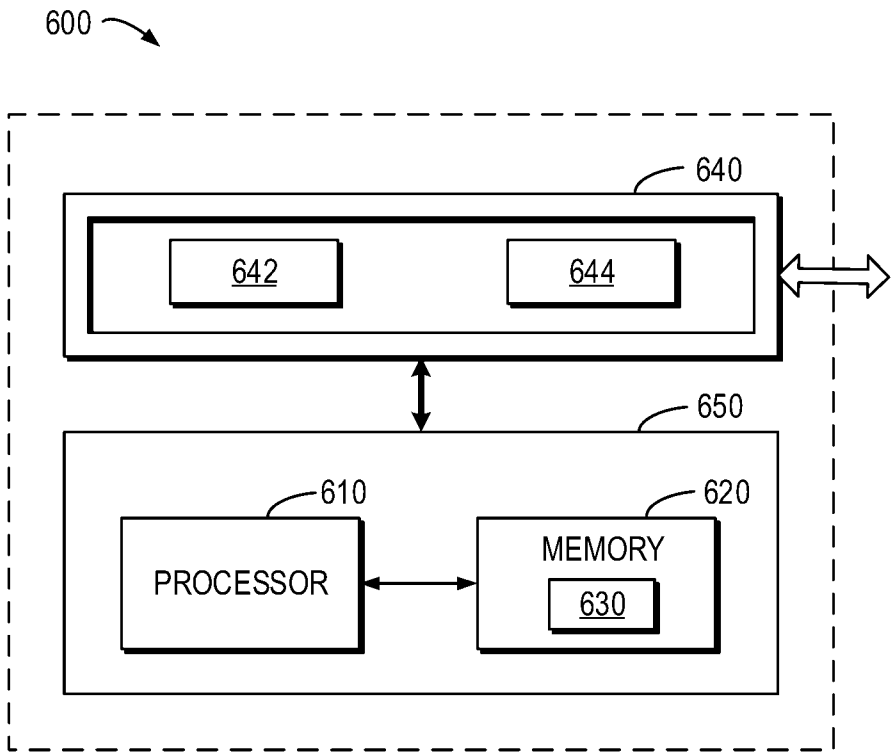

FIG. 6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2023/100799

A. CLASSIFICATION OF SUBJECT MATTER

H04W24/02(2009.01)i

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC: H04L, H04W

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

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

CNTXT, CNKI, WPABS, ENTXTC, 3GPP: availability, timer, machine learning, model, ML, AI, release, identity, active, activate, global model, local model, life cycle management, LCM

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 113765957 A (HUAWEI TECHNOLOGIES CO LTD) 07 December 2021 (2021-12-07) description, paragraphs 0073-0209 and figures 3 to 14	1-20
X	WO 2022077202 A1 (QUALCOMM INC et al.) 21 April 2022 (2022-04-21) description, paragraphs 0068-0140	1-20
A	WO 2021219201 A1 (NOKIA TECHNOLOGIES OY) 04 November 2021 (2021-11-04) the whole document	1-20
A	WO 2022013093 A1 (ERICSSON TELEFON AB L M) 20 January 2022 (2022-01-20) the whole document	1-20
A	WO 2022222089 A1 (QUALCOMM INC et al.) 27 October 2022 (2022-10-27) the whole document	1-20

☐ 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
01 March 2024	**06 March 2024**
Name and mailing address of the ISA/CN **CHINA NATIONAL INTELLECTUAL PROPERTY ADMINISTRATION** **6, Xitucheng Rd., Jimen Bridge, Haidian District, Beijing 100088, China**	Authorized officer **XUE,Yu** Telephone No. **(+86) 010-62412008**

Form PCT/ISA/210 (second sheet) (July 2022)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2023/100799

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	113765957	A	07 December 2021	None			
WO	2022077202	A1	21 April 2022	US	2023379692	A1	23 November 2023
				EP	4229788	A1	23 August 2023
WO	2021219201	A1	04 November 2021	US	2023209384	A1	29 June 2023
				EP	4144120	A1	08 March 2023
WO	2022013093	A1	20 January 2022	US	2023276264	A1	31 August 2023
				EP	4179782	A1	17 May 2023
WO	2022222089	A1	27 October 2022	None			