

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2024 年 7 月 18 日 (18.07.2024)



(10) 国际公布号
WO 2024/149288 A1

- (51) 国际专利分类号:
H04W 16/22 (2009.01)
- (21) 国际申请号: PCT/CN2024/071560
- (22) 国际申请日: 2024 年 1 月 10 日 (10.01.2024)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
202310084335.5 2023年1月12日 (12.01.2023) CN
- (71) 申请人: 维沃移动通信有限公司(VIVO MOBILE COMMUNICATION CO., LTD.) [CN/CN]; 中国广东省东莞市长安镇维沃路 1 号, Guangdong 523863 (CN)。
- (72) 发明人: 孙晓文(SUN, Xiaowen); 中国广东省东莞市长安镇维沃路1号, Guangdong 523863 (CN)。
- (74) 代理人: 北京远志博慧知识产权代理事务所(特殊普通合伙)(BOHUI INTELLECTUAL PROPERTY); 中国北京市海淀区交大东路31号东区10号楼等17幢31幢108, Beijing 100044 (CN)。

- (81) 指定国(除另有指明, 要求每一种可提供的国家保护): 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) 指定国(除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:
— 包括国际检索报告(条约第21条(3))。

(54) Title: AI MODEL DISTRIBUTION METHOD, AI MODEL RECEIVING METHOD, AND TERMINAL AND NETWORK-SIDE DEVICE

(54) 发明名称: AI模型分发、接收方法、终端及网络侧设备

在触发AI模型分发的情况下, 网络侧设备基于至少一个终端的第一信息, 向各个终端发送第一AI模型的信息; 终端的第一信息与终端对应的第一AI模型匹配; 第一信息包括以下至少一项: 终端的能力信息、AI模型的需求信息和传输方式

101

图 2

101 When AI model distribution is triggered, a network-side device sends, on the basis of first information of at least one terminal, information of a first AI model to each terminal, wherein the first information of the terminal matches the first AI model corresponding to the terminal; and the first information comprises at least one of the following: capability information of the terminal, and demand information and a transmission mode of the AI model

(57) Abstract: The present application belongs to the technical field of communications. Disclosed are an AI model distribution method, an AI model receiving method, and a terminal and a network-side device. The AI model distribution method in the embodiments of the present application comprises: when AI model distribution is triggered, a network-side device sending, on the basis of first information of at least one terminal, information of a first AI model to each terminal, wherein the first information of the terminal matches the first AI model corresponding to the terminal; and the first information comprises at least one of the following: capability information of the terminal, and demand information and a transmission mode of the AI model.

(57) 摘要: 本申请公开了一种AI模型分发、接收方法、终端及网络侧设备, 属于通信技术领域, 本申请实施例的AI模型分发方法包括: 在触发AI模型分发的情况下, 网络侧设备基于至少一个终端的第一信息, 向各个所述终端发送第一AI模型的信息; 所述终端的第一信息与所述终端对应的第一AI模型匹配; 所述第一信息包括以下至少一项: 所述终端的能力信息、AI模型的需求信息和传输方式。

WO 2024/149288 A1

AI 模型分发、接收方法、终端及网络侧设备

相关申请的交叉引用

本申请主张在 2023 年 01 月 12 日在中国提交的申请号为 202310084335.5 的中国
5 专利的优先权，其全部内容通过引用包含于此。

技术领域

本申请属于通信技术领域，具体涉及一种 AI 模型分发、接收方法、终端及网络侧设备。

背景技术

10 随着科学技术的发展，机器学习(machine learning, ML)/人工智能(Artificial Intelligence, AI)得到越来越广泛的应用。目前，ML 模型/AI 模型在无线通信网络中承担着越来越多的任务。

例如，在终端进行图像识别或其它媒体任务时，可以通过 AI 模型/ML 模型进行处理，但是，由于终端受到处理能力、存储能力等的限制，因此在处理过程中可能涉及
15 和网络侧设备的交互，例如网络侧设备向终端发送 AI 模型。因此，如何准确地对终端进行 AI 模型分发是需要解决的问题。

发明内容

本申请实施例提供一种 AI 模型分发、接收方法、终端及网络侧设备，能够解决如何准确地对终端进行 AI 模型分发的问题。

20 第一方面，提供了一种 AI 模型分发方法，包括：

在触发 AI 模型分发的情况下，网络侧设备基于至少一个终端的第一信息，向各个所述终端发送第一 AI 模型的信息；所述终端的第一信息与所述终端对应的第一 AI 模型匹配；所述第一信息包括以下至少一项：所述终端的能力信息、AI 模型的需求信息和传输方式。

25 第二方面，提供了一种 AI 模型接收方法，包括：

终端接收网络侧设备发送的第一 AI 模型的信息，所述第一 AI 模型与所述终端的第一信息匹配，所述第一信息包括以下至少一项：所述终端的能力信息、AI 模型的需求信息和传输方式；

所述终端基于所述第一 AI 模型进行业务处理。

30 第三方面，提供了一种 AI 模型分发装置，包括：

发送模块，用于在触发 AI 模型分发的情况下，基于至少一个终端的第一信息，向各个所述终端发送第一 AI 模型的信息；所述终端的第一信息与所述终端对应的第一 AI 模型匹配；所述第一信息包括以下至少一项：所述终端的能力信息、AI 模型的需求

信息和传输方式。

第四方面，提供了一种 AI 模型接收装置，该装置包括：

接收模块，用于接收网络侧设备发送的第一 AI 模型的信息，所述第一 AI 模型与所述终端的第一信息匹配，所述第一信息包括以下至少一项：所述终端的能力信息、

5 AI 模型的需求信息和传输方式；

处理模块，用于基于所述第一 AI 模型进行业务处理。

第五方面，提供了一种终端，该终端包括处理器和存储器，所述存储器存储可在所述处理器上运行的程序或指令，所述程序或指令被所述处理器执行时实现如第二方面所述的方法的步骤。

10 第六方面，提供了一种终端，包括处理器及通信接口；其中，所述通信接口用于接收网络侧设备发送的第一 AI 模型的信息，所述第一 AI 模型与所述终端的第一信息匹配，所述第一信息包括以下至少一项：所述终端的能力信息、AI 模型的需求信息和传输方式；所述处理器用于基于所述第一 AI 模型进行业务处理。

第七方面，提供了一种网络侧设备，该网络侧设备包括处理器和存储器，所述存储器存储可在所述处理器上运行的程序或指令，所述程序或指令被所述处理器执行时实现如第一方面所述的方法的步骤。

第八方面，提供了一种网络侧设备，包括处理器及通信接口；其中，所述通信接口用于在触发 AI 模型分发的情况下，基于至少一个终端的第一信息，向各个所述终端发送第一 AI 模型的信息；所述终端的第一信息与所述终端对应的第一 AI 模型匹配；
20 所述第一信息包括以下至少一项：所述终端的能力信息、AI 模型的需求信息和传输方式。

第九方面，提供了一种通信系统，包括：终端及网络侧设备，所述终端可用于执行如第二方面所述的方法的步骤，所述网络侧设备可用于执行如第一方面所述的方法的步骤。

25 第十方面，提供了一种可读存储介质，所述可读存储介质上存储程序或指令，所述程序或指令被处理器执行时实现如第一方面所述的方法的步骤，或者实现如第二方面所述的方法的步骤。

第十一方面，提供了一种芯片，所述芯片包括处理器和通信接口，所述通信接口和所述处理器耦合，所述处理器用于运行程序或指令，实现如第一方面所述的方法，
30 或实现如第二方面所述的方法。

第十二方面，提供了一种计算机程序/程序产品，所述计算机程序/程序产品被存储在存储介质中，所述计算机程序/程序产品被至少一个处理器执行以实现如第一方面所述的方法的步骤，或者实现如第二方面所述的方法的步骤。

在本申请实施例中，在触发 AI 模型分发的情况下，网络侧设备基于至少一个终端
35 的第一信息，向各个终端发送第一 AI 模型的信息；由于第一 AI 模型为基于终端的第

一信息确定的，例如终端的能力信息、AI模型的需求信息和传输方式中至少一项，即第一信息与终端对应的第一AI模型匹配，使得能够根据实际的终端情况进行AI模型分发，AI模型分发的准确性较高。

附图说明

- 5 图1是本申请实施例可应用的无线通信系统的示意图；
图2是本申请实施例提供的AI模型分发方法的流程示意图；
图3是本申请实施例提供的AI模型分发场景示意图；
图4是本申请实施例提供的AI模型分发方法的交互流程示意图之一；
图5是本申请实施例提供的AI模型分发方法的交互流程示意图之二；
10 图6是本申请实施例提供的AI模型接收方法的流程示意图；
图7是本申请实施例提供的AI模型分发装置的结构示意图；
图8是本申请实施例提供的AI模型接收装置的结构示意图；
图9是本申请实施例提供的通信设备的结构示意图；
图10是本申请实施例提供的终端的结构示意图；
15 图11是本申请实施例的网络侧设备的结构示意图之一；
图12是本申请实施例的网络侧设备的结构示意图之二。

具体实施方式

下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚描述，显然，所描述的实施例是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员所获得的所有其他实施例，都属于本申请保护的范围。

本申请的说明书和权利要求书中的术语“第一”、“第二”等是用于区别类似的对象，而不用来描述特定的顺序或先后次序。应该理解这样使用的术语在适当情况下可以互换，以便本申请的实施例能够以除了在这里图示或描述的那些以外的顺序实施，且“第一”、“第二”所区别的对象通常为一类，并不限定对象的个数，例如第一对象可以是一个，也可以是多个。此外，说明书以及权利要求中“和/或”表示所连接对象的至少其中之一，字符“/”一般表示前后关联对象是一种“或”的关系。本申请的说明书和权利要求书中的术语“指示”既可以是一个明确的指示，也可以是一个隐含的指示。其中，明确的指示可以理解为，发送方在发送的指示中明确告知了接收方需要执行的操作或请求结果；隐含的指示可以理解为，接收方根据发送方发送的指示进行判断，根据判断结果确定需要执行的操作或请求结果。

值得指出的是，本申请实施例所描述的技术不限于长期演进型（Long Term Evolution, LTE）/LTE的演进（LTE-Advanced, LTE-A）系统，还可用于其他无线通信系统，诸如码分多址（Code Division Multiple Access, CDMA）、时分多址（Time Division Multiple Access, TDMA）、频分多址（Frequency Division Multiple Access, FDMA）、正交频分多址（Orthogonal Frequency Division Multiple Access, OFDMA）、

单载波频分多址（Single-carrier Frequency Division Multiple Access, SC-FDMA）和其他系统。本申请实施例中的术语“系统”和“网络”常被可互换地使用，所描述的技术既可用于以上提及的系统 and 无线电技术，也可用于其他系统和无线电技术。以下描述出于示例目的描述了新空口（New Radio, NR）系统，并且在以下大部分描述中使用 NR 术语，但是这些技术也可应用于 NR 系统应用以外的通信系统，如第 6 代（6th Generation, 6G）通信系统。

图 1 是本申请实施例可应用的无线通信系统的示意图，图 1 示出的无线通信系统包括终端 11 和网络侧设备 12。其中，终端 11 可以是手机、平板电脑（Tablet Personal Computer）、膝上型电脑（Laptop Computer）或称为笔记本电脑、个人数字助理（Personal Digital Assistant, PDA）、掌上电脑、上网本、超级移动个人计算机（ultra-mobile personal computer, UMPC）、移动上网装置（Mobile Internet Device, MID）、增强现实（augmented reality, AR）/虚拟现实（virtual reality, VR）设备、机器人、可穿戴式设备（Wearable Device）、车载设备（VUE）、行人终端（PUE）、智能家居（具有无线通信功能的家居设备，如冰箱、电视、洗衣机或者家具等）、游戏机、个人计算机（personal computer, PC）、柜员机或者自助机等终端侧设备，可穿戴式设备包括：智能手表、智能手环、智能耳机、智能眼镜、智能首饰（智能手镯、智能手链、智能戒指、智能项链、智能脚镯、智能脚链等）、智能腕带、智能服装等。除了上述终端设备，也可以是终端内的芯片，例如调制解调器（Modem）芯片，系统级芯片（System on Chip, SoC）。需要说明的是，在本申请实施例并不限定终端 11 的具体类型。

网络侧设备 12 可以包括接入网设备或核心网设备，其中，接入网设备也可以称为无线接入网设备、无线接入网（Radio Access Network, RAN）、无线接入网功能或无线接入网单元。接入网设备可以包括基站、WLAN 接入点或 WiFi 节点等，基站可被称为节点 B、演进节点 B（eNB）、接入点、基收发机站（Base Transceiver Station, BTS）、无线电基站、无线电收发机、基本服务集（Basic Service Set, BSS）、扩展服务集（Extended Service Set, ESS）、家用 B 节点、家用演进型 B 节点、发送接收点（Transmitting Receiving Point, TRP）或所述领域中其他某个合适的术语，只要达到相同的技术效果，所述基站不限于特定技术词汇，需要说明的是，在本申请实施例中仅以 NR 系统中的基站为例进行介绍，并不限定基站的具体类型。核心网设备可以包含但不限于如下至少一项：核心网节点、核心网功能、移动管理实体（Mobility Management Entity, MME）、接入移动管理功能（Access and Mobility Management Function, AMF）、会话管理功能（Session Management Function, SMF）、用户平面功能（User Plane Function, UPF）、策略控制功能（Policy Control Function, PCF）、策略与计费规则功能单元（Policy and Charging Rules Function, PCRF）、边缘应用服务发现功能（Edge Application Server Discovery Function, EASDF）、统一数据管理（Unified Data Management, UDM）、统一数据仓储（Unified Data Repository, UDR）、

归属用户服务器 (Home Subscriber Server, HSS)、集中式网络配置 (Centralized network configuration, CNC)、网络存储功能 (Network Repository Function, NRF)、网络开放功能 (Network Exposure Function, NEF)、本地 NEF (Local NEF, 或 L-NEF)、绑定支持功能 (Binding Support Function, BSF)、应用功能 (Application Function, AF)、位置管理功能 (location manage function, LMF)、增强服务移动定位中心 (Enhanced Serving Mobile Location Centre, E-SMLC)、网络数据分析功能 (network data analytics function, NWDAF) 等。需要说明的是, 在本申请实施例中仅以 NR 系统中的核心网设备为例进行介绍, 并不限定核心网设备的具体类型。

下面结合附图, 通过一些实施例及其应用场景对本申请实施例提供的 AI 模型分发方法、AI 模型接收方法进行详细地说明。

图 2 是本申请实施例提供的 AI 模型分发方法的流程示意图, 如图 2 所示, 该方法包括:

步骤 101、在触发 AI 模型分发的情况下, 网络侧设备基于至少一个终端的第一信息, 向各个终端发送第一 AI 模型的信息; 终端的第一信息与终端对应的第一 AI 模型匹配; 第一信息包括以下至少一项: 终端的能力信息、AI 模型的需求信息和传输方式。

具体地, 考虑到终端能力以及移动性问题, 可能在不同时刻、不同位置, 网络情况、终端的能力 (例如算力) 都是实时变化的, 为了保证分发结果的准确及有效性, 本申请实施例中网络侧设备基于终端的第一信息, 确定终端对应的第一 AI 模型, 并向该终端发送第一 AI 模型的信息。

其中, AI 模型也称为 AI 单元, 可以是 AI 模型、AI 模块或者具备 AI/机器学习 (Machine Learning, ML) 功能的运算规则等。

可选地, 终端的能力信息包括以下至少一项: 计算能力、存储能力、计算能力的使用量、当前剩余计算能力、当前可用计算能力、存储能力的使用量、当前剩余存储能力、当前可用存储能力、可处理的业务信息。

其中, 业务信息例如包括业务标识、业务类型等, 例如计算类型、存储类型。

可选地, AI 模型的需求信息包括以下至少一项: 服务需求信息、AI 模型的描述信息。

可选地, AI 模型的描述信息包括以下至少一项: 标识、功能描述信息、拓扑结构、参数信息、使用的资源信息。

其中传输方式例如包括: 通过 5G 系统 (5G system, 5GS) 网络, 或直接通过应用层传输 (即 5GS 不感知 (over 5GS))。

进一步, 终端接收到第一 AI 模型的信息, 可以基于第一 AI 模型进行业务处理, 例如图像识别。

其中, 可以通过终端和网络侧设备的通信触发 AI 模型的传递, 例如终端发送请求。

本申请实施例提供的 AI 模型分发方法中, 在触发 AI 模型分发的情况下, 网络侧

设备基于至少一个终端的第一信息，向各个终端发送第一 AI 模型的信息；由于第一 AI 模型为基于终端的第一信息确定的，例如终端的能力信息、AI 模型的需求信息和传输方式中至少一项，即第一信息与终端对应的第一 AI 模型匹配，使得能够根据实际的终端情况进行 AI 模型分发，AI 模型分发的准确性较高。

5 可选地，步骤 101 之前还包括：

网络侧设备接收至少一个终端的请求消息，请求消息用于请求网络侧设备发送第一 AI 模型的信息，请求消息包括：终端的第一信息。

具体地，在触发 AI 模型传递之前，可以在网络侧设备和终端之间交换和协商与所需 AI 模型相关的信息。例如包括终端的能力信息（包括但不限于计算能力、存储能力、
10 当前计算能力的使用量等）、AI 模型的需求信息（例如，服务需求信息、AI 模型的描述信息，服务需求信息例如所需 AI 模型可处理的业务信息、类型等）和传输方式相关的信息。上述信息可以用于为终端所需服务选择合适的 AI 模型。

可选地，步骤 101 之前，还包括：

网络侧设备建立与终端的第一会话连接；第一会话连接用于传输第一 AI 模型的信息。
15

可选地，该方法还包括：

网络侧设备接收至少一个终端发送的训练结果数据，训练结果数据为终端对第一 AI 模型进行训练得到的；

网络侧设备基于至少一个终端发送的训练结果数据进行 AI 模型训练。

具体地，终端接收到第一 AI 模型的信息之后，可以基于训练数据对第一 AI 模型进行训练，训练数据例如为终端获取的训练样本数据。进一步，可以将训练得到的训练结果数据发送给网络侧设备，网络侧设备可以基于训练结果数据进行 AI 模型训练，训练结果数据例如可以包括训练后的 AI 模型的描述信息，例如包括：标识、功能描述信息、拓扑结构、参数信息、使用的资源信息等。
20

25 可选地，该方法还包括：

网络侧设备建立与终端的第二会话连接，第二会话连接用于传输训练结果数据。

上述实施方式中，通过终端对 AI 模型进行训练，将训练结果数据反馈给网络侧设备，使得网络侧设备可以基于终端的反馈进一步修正模型，进而能够满足终端的需求。

在一实施例中，在触发 AI 模型更新的情况下，网络侧设备基于第二信息向至少一个终端发送更新后的第二 AI 模型的信息；第二信息包括以下至少一项：各个终端的能力更新信息、模型更新信息。
30

具体地，网络侧设备基于第二信息向至少一个终端发送更新后的第二模型的信息，例如基于终端最新的能力信息、最新的服务需求信息、模型更新信息等选择更新的与该终端匹配的 AI 模型，模型更新信息例如为网络侧设备基于至少一个终端的训练结果数据更新的新的任务（如联邦学习任务）训练得到的 AI 模型。
35

可选地，能力更新信息包括以下至少一项的更新信息：计算能力、存储能力、计算能力的使用量、当前剩余计算能力、当前可用计算能力、存储能力的使用量、当前剩余存储能力、当前可用存储能力、可处理的业务信息；

5 可选地，模型更新信息包括以下至少一项的更新信息：标识、功能描述信息、拓扑结构、参数信息、使用的资源信息。

功能描述信息例如用于对 AI 模型的功能进行说明的信息，AI 模型的功能例如可以处理的业务名称、业务类型等；拓扑结构是指 AI 模型采用的网络模型的结构，例如神经网络模型的拓扑结构、深度学习模型的拓扑结构等；参数信息例如包括 AI 模型的权重信息、采用的损失函数信息、激活函数、卷积核大小、卷积层数量等；使用的资源信息例如包括：运行时占用的内存大小、存储空间大小等。

10

例如，模型的标识进行了更新，则模型更新信息包括标识的更新信息，例如包括更新前的标识以及更新后的标识，或，模型的功能进行了更新，则模型更新信息包括功能描述信息的更新信息，例如包括更新后的功能描述信息，拓扑结构、参数信息、使用的资源信息等类似。

15 可选地，第二 AI 模型的信息包括以下至少一项：标识、功能描述信息、拓扑结构、参数信息、使用的资源信息。

可选地，第二 AI 模型为基于至少一个终端的训练结果数据对第一 AI 模型进行训练得到的；训练结果数据为终端对所述第一 AI 模型进行训练得到的。

可选地，AI 模型更新为基于第三信息触发的，第三信息包括以下至少一项：终端的能力更新指示信息、网络侧设备更新 AI 模型的指示信息；终端的能力更新指示信息用于指示以下至少一项：终端的计算能力使用率大于或等于第一阈值、终端的计算资源使用率大于或等于第二阈值、终端的负载大于或等于第三阈值、终端的资源小于或等于第四阈值。

20

具体地，AI 模型更新的触发因素可以是以下至少一项：

25 (1) 终端当前的计算能力改变（例如内存、硬盘等使用率发生改变），计算能力不能满足当前 AI 模型的需求或远超于当前 AI 模型的需求；

(2) 网络侧设备更新 AI 模型。

上述实施方式中，在触发 AI 模型更新的情况下，网络侧设备基于至少一个终端的第二信息，向各个终端发送更新的第二 AI 模型的信息；由于第二 AI 模型为基于终端的第二信息确定的，例如终端的能力更新信息、AI 模型的需求更新信息、服务需求更新信息中至少一项，使得能够根据实际的终端情况进行 AI 模型分发，AI 模型分发的准确性较高。

30

示例性地，如图 3 所示的架构中，对于网络侧来说，联邦学习引擎（Federated learning engine）从 AI 模型库（AI Model repository）接收部分训练模型，该模型被传递到 AI 模型交付功能（AI model delivery function），以通过 5GS 交付给多个终端。

35

来自多个终端的训练结果数据(Training Result Data)也通过 5GS 被联邦学习引擎接收, 然后聚合用于 AI 模型的持续训练。

AI 模型的更新(例如在拓扑、权重等参数或其他信息)在更新过程中被传送到终端。

5 对于终端侧来说, AI 模型数据通过 5GS 由 AI 模型访问功能(AI model access function)接收, 然后将数据传递给 AI 训练引擎(AI training engine)。

例如, 终端中的 AI 训练引擎使用本地设备数据作为输入的训练数据来训练 AI 模型。

10 训练结果(例如以更新的 AI 模型的描述信息的形式)通过训练结果传递功能(Training results delivery function)传递到网络侧。

需要说明的是, 网络侧设备可以是接入网设备和/或核心网设备, 上述功能模块可以部署在一个或多个网络侧设备中。

示例性地, 如图 4 所示, 该方法包括:

- 步骤 1、终端 APP 和网络 APP 通信触发 AI 模型的传递;
- 15 步骤 2、终端 APP 和网络 APP 交互, 并选择 AI 模型;
- 步骤 3、网络 APP 标识被选中的 AI 模型
- 步骤 4、AI 模型交付功能与 AI 模型访问功能建立用于 AI 模型传递的会话连接;
- 步骤 5、AI 模型访问功能接收 AI 模型;
- 步骤 6、AI 模型访问功能将 AI 模型传递至 AI 训练引擎;
- 20 步骤 7、终端上, 训练数据传递至 AI 训练引擎;
- 步骤 8、AI 训练引擎执行模型训练;
- 步骤 9、训练结果传递功能与联邦学习引擎建立用于训练结果传递的会话连接;
- 步骤 10、训练结果传递至联邦学习引擎。

示例性地, 如图 5 所示, 该方法包括:

- 25 步骤 0、触发 AI 模型更新, 触发因素例如包括但不限于以下几条:
 - (1) 终端当前的计算能力改变(例如内存、硬盘等使用率发生改变);
 - (2) 联邦学习引擎更新 AI 模型。
- 步骤 1、终端 APP 和网络 APP 通信触发 AI 模型的更新传递;
- 步骤 2、终端 APP 和网络 APP 交互, 并选择更新的 AI 模型;
- 30 步骤 3、网络 APP 标识被选中的更新的 AI 模型
- 步骤 4、AI 模型交付功能与 AI 模型访问功能建立用于 AI 模型传递的会话连接;
- 步骤 5、AI 模型访问功能接收更新的 AI 模型;
- 步骤 6、AI 模型访问功能将更新的 AI 模型传递至 AI 训练引擎;
- 步骤 7、终端上, training input data 传递至 AI training engine
- 35 步骤 8、AI 训练引擎执行模型训练;

步骤 9、训练结果传递功能与联邦学习引擎建立用于训练结果传递的会话连接；
步骤 10、训练结果传递至联邦学习引擎。

图 6 是本申请实施例提供的 AI 模型接收方法的流程示意图，如图 6 所示，该方法包括：

5 步骤 201、终端接收网络侧设备发送的第一 AI 模型的信息，第一 AI 模型与终端的第一信息匹配，第一信息包括以下至少一项：终端的能力信息、AI 模型的需求信息和传输方式；

步骤 202、终端基于第一 AI 模型进行业务处理。

可选地，所述方法还包括：

10 所述终端基于训练数据对所述第一 AI 模型进行训练，得到训练结果数据；

所述终端向所述网络侧设备发送所述训练结果数据；所述训练结果数据用于所述网络侧设备进行 AI 模型训练。

可选地，所述终端的能力信息包括以下至少一项：计算能力、存储能力、计算能力的使用量、当前剩余计算能力、当前可用计算能力、存储能力的使用量、当前剩余
15 存储能力、当前可用存储能力、可处理的业务信息。

可选地，所述 AI 模型的需求信息包括以下至少一项：服务需求信息、AI 模型的描述信息；所述 AI 模型的描述信息包括：标识、功能描述信息、拓扑结构、参数信息、使用的资源信息。

可选地，所述方法还包括：

20 所述终端接收所述网络侧设备发送的更新后的第二 AI 模型的信息，所述第二 AI 模型为基于第二信息得到的；所述第二信息包括以下至少一项：各个所述终端的能力更新信息、服务需求更新信息、模型更新信息。

可选地，所述更新后的第二 AI 模型的信息为触发 AI 模型更新的情况下发送的，所述 AI 模型更新为基于第三信息触发的，所述第三信息包括以下至少一项：所述终端
25 的能力更新指示信息、所述网络侧设备更新 AI 模型的指示信息；所述终端的能力更新指示信息用于指示以下至少一项：所述终端的计算能力使用率大于或等于第一阈值、所述终端的计算资源使用率大于或等于第二阈值、所述终端的负载大于或等于第三阈值、所述终端的资源小于或等于第四阈值。

可选地，所述第二 AI 模型为基于至少一个所述终端的训练结果数据对所述第一
30 AI 模型进行训练得到的；所述训练结果数据为所述终端对所述第一 AI 模型进行训练得到的。

可选地，所述第二 AI 模型的信息包括以下至少一项：标识、功能描述信息、拓扑结构、参数信息、使用的资源信息。

可选地，所述训练结果数据包括更新后的 AI 模型的描述信息。

35 可选地，所述终端接收网络侧设备发送的第一 AI 模型的信息之前，还包括：

所述终端向所述网络侧设备发送请求消息，所述请求消息用于请求所述网络侧设备发送所述第一 AI 模型的信息，所述请求消息包括：所述终端的第一信息。

可选地，所述终端接收网络侧设备发送的第一 AI 模型的信息之前，还包括：

所述终端建立与所述网络侧设备的第一会话连接；所述第一会话连接用于传输第一 AI 模型的信息。

可选地，所述方法还包括：

所述终端建立与所述网络侧设备的第二会话连接，所述第二会话连接用于传输训练结果数据。

本实施例的方法，其具体实现过程与技术效果与终端侧方法实施例中相同，具体可以参见终端侧方法实施例中的详细介绍，此处不再赘述。

本申请实施例提供的 AI 模型分发方法，执行主体可以为 AI 模型分发装置。本申请实施例提供的 AI 模型接收方法，执行主体可以为 AI 模型接收装置。本申请实施例中以 AI 模型分发装置执行 AI 模型分发方法、AI 模型接收装置执行 AI 模型接收方法为例，说明本申请实施例提供的 AI 模型分发装置、AI 模型接收装置。

图 7 是本申请实施例提供的 AI 模型分发装置的结构示意图，如图 7 所示，该 AI 模型分发装置，应用于网络侧设备，包括：

发送模块 210，用于在触发 AI 模型分发的情况下，基于至少一个终端的第一信息，向各个所述终端发送第一 AI 模型的信息；所述终端的第一信息与所述终端对应的第一 AI 模型匹配；所述第一信息包括以下至少一项：所述终端的能力信息、AI 模型的需求信息和传输方式。

可选地，所述终端的能力信息包括以下至少一项：计算能力、存储能力、计算能力的使用量、当前剩余计算能力、当前可用计算能力、存储能力的使用量、当前剩余存储能力、当前可用存储能力、可处理的业务信息；

可选地，所述 AI 模型的需求信息包括以下至少一项：服务需求信息、AI 模型的描述信息。

可选地，所述 AI 模型的描述信息包括以下至少一项：标识、功能描述信息、拓扑结构、参数信息、使用的资源信息。

可选地，该装置还包括：

接收模块，用于接收至少一个所述终端发送的训练结果数据，所述训练结果数据为所述终端对所述第一 AI 模型进行训练得到的；

处理模块，用于基于至少一个所述终端发送的训练结果数据进行 AI 模型训练。

可选地，发送模块 210，还用于在触发 AI 模型更新的情况下，基于第二信息向至少一个所述终端发送更新后的第二 AI 模型的信息；所述第二信息包括以下至少一项：各个所述终端的能力更新信息、服务需求更新信息、模型更新信息。

可选地，所述能力更新信息包括以下至少一项的更新信息：计算能力、存储能力、

计算能力的使用量、当前剩余计算能力、当前可用计算能力、存储能力的使用量、当前剩余存储能力、当前可用存储能力、可处理的业务信息；和/或，

所述模型更新信息包括以下至少一项的更新信息：标识、功能描述信息、拓扑结构、参数信息、使用的资源信息。

5 可选地，所述 AI 模型更新为基于第三信息触发的，所述第三信息包括以下至少一项：终端的能力更新指示信息、所述网络侧设备更新 AI 模型的指示信息；所述终端的能力更新指示信息用于指示以下至少一项：所述终端的计算能力使用率大于或等于第一阈值、所述终端的计算资源使用率大于或等于第二阈值、所述终端的负载大于或等于第三阈值、所述终端的资源小于或等于第四阈值。

10 可选地，所述第二 AI 模型为基于至少一个所述终端的训练结果数据对所述第一 AI 模型进行训练得到的；所述训练结果数据为所述终端对所述第一 AI 模型进行训练得到的。

可选地，所述第二 AI 模型的信息包括以下至少一项：标识、功能描述信息、拓扑结构、参数信息、使用的资源信息。

15 可选地，所述训练结果数据包括更新后的 AI 模型的描述信息。

可选地，接收模块，还用于接收至少一个所述终端的请求消息，所述请求消息用于请求所述网络侧设备发送所述第一 AI 模型的信息，所述请求消息包括：所述终端的第一信息。

可选地，处理模块，还用于：

20 建立与所述终端的第一会话连接；所述第一会话连接用于传输第一 AI 模型的信息。

可选地，处理模块，还用于：

建立与所述终端的第二会话连接，所述第二会话连接用于传输训练结果数据。

25 本实施例的装置，可以用于执行前述网络侧设备方法实施例中任一实施例的方法，其具体实现过程与技术效果与网络侧设备方法实施例中相同，具体可以参见网络侧设备方法实施例中的详细介绍，此处不再赘述。

图 8 是本申请实施例提供的 AI 模型接收装置的结构示意图，如图 8 所示，该 AI 模型接收装置，应用于终端，包括：

30 接收模块 310，用于接收网络侧设备发送的第一 AI 模型的信息，所述第一 AI 模型与所述终端的第一信息匹配，所述第一信息包括以下至少一项：所述终端的能力信息、AI 模型的需求信息和传输方式；

处理模块 320，用于基于所述第一 AI 模型进行业务处理。

可选地，所述处理模块 320，还用于：

基于训练数据对所述第一 AI 模型进行训练，得到训练结果数据；

35 该装置还包括：发送模块，用于向所述网络侧设备发送所述训练结果数据；所述

训练结果数据用于所述网络侧设备进行 AI 模型训练。

可选地，所述终端的能力信息包括以下至少一项：计算能力、存储能力、计算能力的使用量、当前剩余计算能力、当前可用计算能力、存储能力的使用量、当前剩余存储能力、当前可用存储能力、可处理的业务信息。

5 可选地，所述 AI 模型的需求信息包括以下至少一项：服务需求信息、AI 模型的描述信息；所述 AI 模型的描述信息包括：标识、功能描述信息、拓扑结构、参数信息、使用的资源信息。

可选地，接收模块 310，还用于：

10 接收所述网络侧设备发送的更新后的第二 AI 模型的信息，所述第二 AI 模型为基于第二信息得到的；所述第二信息包括以下至少一项：各个所述终端的能力更新信息、服务需求更新信息、模型更新信息。

可选地，所述能力更新信息包括以下至少一项的更新信息：计算能力、存储能力、计算能力的使用量、当前剩余计算能力、当前可用计算能力、存储能力的使用量、当前剩余存储能力、当前可用存储能力、可处理的业务信息；和/或

15 所述模型更新信息包括以下至少一项的更新信息：标识、功能描述信息、拓扑结构、参数信息、使用的资源信息。

可选地，所述更新后的第二 AI 模型的信息为触发 AI 模型更新的情况下发送的，所述 AI 模型更新为基于第三信息触发的，所述第三信息包括以下至少一项：所述终端的能力更新指示信息、所述网络侧设备更新 AI 模型的指示信息；所述终端的能力更新指示信息用于指示以下至少一项：所述终端的计算能力使用率大于或等于第一阈值、所述终端的计算资源使用率大于或等于第二阈值、所述终端的负载大于或等于第三阈值、所述终端的资源小于或等于第四阈值。

20 可选地，所述第二 AI 模型为基于至少一个所述终端的训练结果数据对所述第一 AI 模型进行训练得到的；所述训练结果数据为所述终端对所述第一 AI 模型进行训练得到的。

可选地，所述第二 AI 模型的信息包括以下至少一项：标识、功能描述信息、拓扑结构、参数信息、使用的资源信息。

可选地，所述训练结果数据包括更新后的 AI 模型的描述信息。

可选地，发送模块，还用于：

30 向所述网络侧设备发送请求消息，所述请求消息用于请求所述网络侧设备发送所述第一 AI 模型的信息，所述请求消息包括：所述终端的第一信息。

可选地，处理模块 320，还用于建立与所述网络侧设备的第一会话连接；所述第一会话连接用于传输第一 AI 模型的信息。

35 可选地，处理模块 320，还用于建立与所述网络侧设备的第二会话连接，所述第二会话连接用于传输训练结果数据。

本实施例的装置，可以用于执行前述终端侧方法实施例中任一实施例的方法，其具体实现过程与技术效果与终端侧方法实施例中相同，具体可以参见终端侧方法实施例中的详细介绍，此处不再赘述。

本申请实施例中的 AI 模型分发装置、AI 模型接收装置可以是电子设备，例如具有操作系统的电子设备，也可以是电子设备中的部件，例如集成电路或芯片。该电子设备可以是终端，也可以为除终端之外的其他设备。示例性的，终端可以包括但不限于上述所列举的终端 11 的类型，其他设备可以为服务器、网络附属存储器（Network Attached Storage，NAS）等，本申请实施例不作具体限定。

本申请实施例提供的 AI 模型分发装置、AI 模型接收装置能够实现图 2 至图 6 的方法实施例实现的各个过程，并达到相同的技术效果，为避免重复，这里不再赘述。

图 9 是本申请实施例提供的通信设备的结构示意图，如图 9 所示，该通信设备 900，包括处理器 901 和存储器 902，存储器 902 上存储有可在所述处理器 901 上运行的程序或指令，例如，该通信设备 900 为终端时，该程序或指令被处理器 901 执行时实现上述 AI 模型接收方法实施例的各个步骤，且能达到相同的技术效果。该通信设备 900 为网络侧设备时，该程序或指令被处理器 901 执行时实现上述 AI 模型分发方法实施例的各个步骤，且能达到相同的技术效果，为避免重复，这里不再赘述。

本申请实施例还提供一种终端，包括处理器和通信接口，通信接口用于接收网络侧设备发送的第一 AI 模型的信息，所述第一 AI 模型与所述终端的第一信息匹配，所述第一信息包括以下至少一项：所述终端的能力信息、AI 模型的需求信息和传输方式；处理器用于基于所述第一 AI 模型进行业务处理。该终端实施例与上述终端侧方法实施例对应，上述方法实施例的各个实施过程和实现方式均可适用于该终端实施例中，且能达到相同的技术效果。

图 10 是本申请实施例提供的终端的结构示意图，如图 10 所示，该终端 1000 包括但不限于：射频单元 1001、网络模块 1002、音频输出单元 1003、输入单元 1004、传感器 1005、显示单元 1006、用户输入单元 1007、接口单元 1008、存储器 1009、以及处理器 1010 等中的至少部分部件。

本领域技术人员可以理解，终端 1000 还可以包括给各个部件供电的电源（比如电池），电源可以通过电源管理系统与处理器 1010 逻辑相连，从而通过电源管理系统实现管理充电、放电、以及功耗管理等功能。图 10 中示出的终端结构并不构成对终端的限定，终端可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件布置，在此不再赘述。

应理解的是，本申请实施例中，输入单元 1004 可以包括图形处理单元（Graphics Processing Unit，GPU）10041 和麦克风 10042，图形处理器 10041 对在视频捕获模式或图像捕获模式中由图像捕获装置（如摄像头）获得的静态图片或视频的图像数据进行处理。显示单元 1006 可包括显示面板 10061，可以采用液晶显示器、有机发光二极

管等形式来配置显示面板 10061。用户输入单元 1007 包括触控面板 10071 以及其它输入设备 10072 中的至少一种。触控面板 10071，也称为触摸屏。触控面板 10071 可包括触摸检测装置和触摸控制器两个部分。其它输入设备 10072 可以包括但不限于物理键盘、功能键（比如音量控制按键、开关按键等）、轨迹球、鼠标、操作杆，在此不再赘述。

本申请实施例中，射频单元 1001 将接收来自网络侧设备的下行数据接收后，可以传输给处理器 1010 进行处理；另外，射频单元 1001 可以将上行的数据发送给向网络侧设备发送上行数据。通常，射频单元 1001 包括但不限于天线、至少一个放大器、收发信机、耦合器、低噪声放大器、双工器等。

存储器 1009 可用于存储软件程序或指令以及各种数据。存储器 1009 可主要包括存储程序或指令的第一存储区和存储数据的第二存储区，其中，第一存储区可存储操作系统、至少一个功能所需的应用程序或指令（比如声音播放功能、图像播放功能等）等。此外，存储器 1009 可以包括易失性存储器或非易失性存储器，或者，存储器 1009 可以包括瞬态和非瞬态存储器。其中，非易失性存储器可以是只读存储器(Read-Only Memory, ROM)、可编程只读存储器(Programmable ROM, PROM)、可擦除可编程只读存储器(Erasable PROM, EPROM)、电可擦除可编程只读存储器(Electrically EPROM, EEPROM)或闪存。易失性存储器可以是随机存取存储器(Random Access Memory, RAM)、静态随机存取存储器(Static RAM, SRAM)、动态随机存取存储器(Dynamic RAM, DRAM)、同步动态随机存取存储器(Synchronous DRAM, SDRAM)、双倍数据速率同步动态随机存取存储器(Double Data Rate SDRAM, DDRSDRAM)、增强型同步动态随机存取存储器(Enhanced SDRAM, ESDRAM)、同步连接动态随机存取存储器(Synch link DRAM, SLDRAM)和直接内存总线随机存取存储器(Direct Rambus RAM, DRRAM)。本申请实施例中的存储器 1009 包括但不限于这些和任意其它适合类型的存储器。

处理器 1010 可包括一个或多个处理单元；可选的，处理器 1010 可集成应用处理器和调制解调处理器，其中，应用处理器主要处理涉及操作系统、用户界面和应用程序或指令等的操作，调制解调处理器主要处理无线通信信号，如基带处理器。可以理解的是，上述调制解调处理器也可以不集成到处理器 1010 中。

其中，射频单元 1001，用于接收网络侧设备发送的第一 AI 模型的信息，所述第一 AI 模型与所述终端的第一信息匹配，所述第一信息包括以下至少一项：所述终端的能力信息、AI 模型的需求信息和传输方式；

处理器 1010，用于基于所述第一 AI 模型进行业务处理。

可选地，所述处理器 1010，还用于：

基于训练数据对所述第一 AI 模型进行训练，得到训练结果数据；

射频单元 1001，还用于向所述网络侧设备发送所述训练结果数据；所述训练结果

数据用于所述网络侧设备进行 AI 模型训练。

可选地，所述终端的能力信息包括以下至少一项：计算能力、存储能力、计算能力的使用量、当前剩余计算能力、当前可用计算能力、存储能力的使用量、当前剩余存储能力、当前可用存储能力、可处理的业务信息。

5 可选地，所述 AI 模型的需求信息包括以下至少一项：服务需求信息、AI 模型的描述信息；所述 AI 模型的描述信息包括：标识、功能描述信息、拓扑结构、参数信息、使用的资源信息。

可选地，射频单元 1001，还用于：

10 接收所述网络侧设备发送的更新后的第二 AI 模型的信息，所述第二 AI 模型为基于第二信息得到的；所述第二信息包括以下至少一项：各个所述终端的能力更新信息、服务需求更新信息、模型更新信息。

可选地，所述能力更新信息包括以下至少一项的更新信息：计算能力、存储能力、计算能力的使用量、当前剩余计算能力、当前可用计算能力、存储能力的使用量、当前剩余存储能力、当前可用存储能力、可处理的业务信息；和/或

15 所述模型更新信息包括以下至少一项的更新信息：标识、功能描述信息、拓扑结构、参数信息、使用的资源信息。

可选地，所述更新后的第二 AI 模型的信息为触发 AI 模型更新的情况下发送的，所述 AI 模型更新为基于第三信息触发的，所述第三信息包括以下至少一项：所述终端的能力更新指示信息、所述网络侧设备更新 AI 模型的指示信息；所述终端的能力更新指示信息用于指示以下至少一项：所述终端的计算能力使用率大于或等于第一阈值、所述终端的计算资源使用率大于或等于第二阈值、所述终端的负载大于或等于第三阈值、所述终端的资源小于或等于第四阈值。

20 可选地，所述第二 AI 模型为基于至少一个所述终端的训练结果数据对所述第一 AI 模型进行训练得到的；所述训练结果数据为所述终端对所述第一 AI 模型进行训练得到的。

可选地，所述第二 AI 模型的信息包括以下至少一项：标识、功能描述信息、拓扑结构、参数信息、使用的资源信息。

可选地，所述训练结果数据包括更新后的 AI 模型的描述信息。

可选地，射频单元 1001，还用于：

30 向所述网络侧设备发送请求消息，所述请求消息用于请求所述网络侧设备发送所述第一 AI 模型的信息，所述请求消息包括：所述终端的第一信息。

可选地，处理器 1010，还用于建立与所述网络侧设备的第一会话连接；所述第一会话连接用于传输第一 AI 模型的信息。

35 可选地，处理器 1010，还用于建立与所述网络侧设备的第二会话连接，所述第二会话连接用于传输训练结果数据。

本申请实施例还提供一种网络侧设备，包括处理器和通信接口，通信接口用于在触发 AI 模型分发的情况下，基于至少一个终端的第一信息，向各个所述终端发送第一 AI 模型的信息；所述终端的第一信息与所述终端对应的第一 AI 模型匹配；所述第一信息包括以下至少一项：所述终端的能力信息、AI 模型的需求信息和传输方式。该网络侧设备实施例与上述网络侧设备方法实施例对应，上述方法实施例的各个实施过程和实现方式均可适用于该网络侧设备实施例中，且能达到相同的技术效果。

图 11 是本申请实施例提供的网络侧设备的结构示意图，如图 11 所示，该网络侧设备 1100 包括：天线 1101、射频装置 1102、基带装置 1103、处理器 1104 和存储器 1105。天线 1101 与射频装置 1102 连接。在上行方向上，射频装置 1102 通过天线 1101 接收信息，将接收的信息发送给基带装置 1103 进行处理。在下行方向上，基带装置 1103 对要发送的信息进行处理，并发送给射频装置 1102，射频装置 1102 对收到的信息进行处理后经过天线 1101 发送出去。

以上实施例中网络侧设备执行的方法可以在基带装置 1103 中实现，该基带装置 1103 包括基带处理器。

基带装置 1103 例如可以包括至少一个基带板，该基带板上设置有多个芯片，如图 11 所示，其中一个芯片例如为基带处理器，通过总线接口与存储器 1105 连接，以调用存储器 1105 中的程序，执行以上方法实施例中所示的网络设备操作。

该网络侧设备还可以包括网络接口 1106，该接口例如为通用公共无线接口（common public radio interface，CPRI）。

具体地，本申请实施例的网络侧设备 1100 还包括：存储在存储器 1105 上并可在处理器 1104 上运行的指令或程序，处理器 1104 调用存储器 1105 中的指令或程序执行如上所述的 AI 模型的分发方法，并达到相同的技术效果，为避免重复，故不在此赘述。

图 12 是本申请实施例提供的网络侧设备的结构示意图，如图 12 所示，该网络侧设备 1200 包括：处理器 1201、网络接口 1202 和存储器 1203。其中，网络接口 1202 例如为通用公共无线接口（common public radio interface，CPRI）。

具体地，本申请实施例的网络侧设备 1200 还包括：存储在存储器 1203 上并可在处理器 1201 上运行的指令或程序，处理器 1201 调用存储器 1203 中的指令或程序执行图 7 所示各模块执行的方法，并达到相同的技术效果，为避免重复，故不在此赘述。

本申请实施例还提供了一种通信系统，包括：终端及网络侧设备，所述终端可用于执行如上所述的 AI 模型接收方法的步骤，所述网络侧设备可用于执行如上所述的 AI 模型分发方法的步骤。

本申请实施例还提供一种可读存储介质，所述可读存储介质可以是以易失性的，也可以是非易失性的，所述可读存储介质上存储有程序或指令，该程序或指令被处理器执行时实现上述 AI 模型分发、AI 模型接收方法实施例的各个过程，且能达到相同的技术效果，为避免重复，这里不再赘述。

其中，所述处理器为上述实施例中所述的终端中的处理器。所述可读存储介质，包括计算机可读存储介质，如计算机只读存储器 ROM、随机存取存储器 RAM、磁碟或者光盘等。

本申请实施例另提供了一种芯片，所述芯片包括处理器和通信接口，所述通信接口和所述处理器耦合，所述处理器用于运行程序或指令，实现上述 AI 模型分发、AI 模型接收方法实施例的各个过程，且能达到相同的技术效果，为避免重复，这里不再赘述。

应理解，本申请实施例提到的芯片还可以称为系统级芯片，系统芯片，芯片系统或片上系统芯片等。

本申请实施例另提供了一种计算机程序/程序产品，所述计算机程序/程序产品被存储在存储介质中，所述计算机程序/程序产品被至少一个处理器执行以实现上述 AI 模型分发、AI 模型接收方法实施例的各个过程，且能达到相同的技术效果，为避免重复，这里不再赘述。

需要说明的是，在本文中，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者装置不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者装置所固有的要素。在没有更多限制的情况下，由语句“包括一个...”限定的要素，并不排除在包括该要素的过程、方法、物品或者装置中还存在另外的相同要素。此外，需要指出的是，本申请实施方式中的方法和装置的范围不限按示出或讨论的顺序来执行功能，还可包括根据所涉及的功能按基本同时的方式或按相反的顺序来执行功能，例如，可以按不同于所描述的次序来执行所描述的方法，并且还可以添加、省去、或组合各种步骤。另外，参照某些示例所描述的特征可在其他示例中被组合。

通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件，但很多情况下前者是更佳的实施方式。基于这样的理解，本申请的技术方案本质上或者说对相关技术做出贡献的部分可以以计算机软件产品的形式体现出来，该计算机软件产品存储在一个存储介质（如 ROM/RAM、磁碟、光盘）中，包括若干指令用以使得一台终端（可以是手机，计算机，服务器，空调器，或者网络设备等）执行本申请各个实施例所述的方法。

上面结合附图对本申请的实施例进行了描述，但是本申请并不局限于上述的具体实施方式，上述的具体实施方式仅仅是示意性的，而不是限制性的，本领域的普通技术人员在本申请的启示下，在不脱离本申请宗旨和权利要求所保护的范围情况下，还可做出很多形式，均属于本申请的保护之内。

权 利 要 求 书

1. 一种人工智能 AI 模型分发方法，包括：

在触发 AI 模型分发的情况下，网络侧设备基于至少一个终端的第一信息，向各个所述终端发送第一 AI 模型的信息；所述终端的第一信息与所述终端对应的第一 AI 模型匹配；所述第一信息包括以下至少一项：所述终端的能力信息、AI 模型的需求信息和传输方式。

2. 根据权利要求 1 所述的 AI 模型分发方法，其中，

所述终端的能力信息包括以下至少一项：计算能力、存储能力、计算能力的使用量、当前剩余计算能力、当前可用计算能力、存储能力的使用量、当前剩余存储能力、当前可用存储能力、可处理的业务信息。

3. 根据权利要求 1 所述的 AI 模型分发方法，其中，

所述 AI 模型的需求信息包括以下至少一项：服务需求信息、AI 模型的描述信息。

4. 根据权利要求 3 所述的 AI 模型分发方法，其中，

所述 AI 模型的描述信息包括以下至少一项：标识、功能描述信息、拓扑结构、参数信息、使用的资源信息。

5. 根据权利要求 1 或 2 所述的 AI 模型分发方法，其中，所述方法还包括：

所述网络侧设备接收至少一个所述终端发送的训练结果数据，所述训练结果数据为所述终端对所述第一 AI 模型进行训练得到的；

所述网络侧设备基于至少一个所述终端发送的训练结果数据进行 AI 模型训练。

6. 根据权利要求 1 或 2 所述的 AI 模型分发方法，其中，所述方法还包括：

在触发 AI 模型更新的情况下，所述网络侧设备基于第二信息向至少一个所述终端发送更新后的第二 AI 模型的信息；所述第二信息包括以下至少一项：各个所述终端的能力更新信息、服务需求更新信息、模型更新信息。

7. 根据权利要求 6 所述的 AI 模型分发方法，其中，

所述能力更新信息包括以下至少一项的更新信息：计算能力、存储能力、计算能力的使用量、当前剩余计算能力、当前可用计算能力、存储能力的使用量、当前剩余存储能力、当前可用存储能力、可处理的业务信息；和/或，

所述模型更新信息包括以下至少一项的更新信息：标识、功能描述信息、拓扑结构、参数信息、使用的资源信息。

8. 根据权利要求 6 所述的 AI 模型分发方法，其中，

所述 AI 模型更新为基于第三信息触发的，所述第三信息包括以下至少一项：终端的能力更新指示信息、所述网络侧设备更新 AI 模型的指示信息；所述终端的能力更新指示信息用于指示以下至少一项：所述终端的计算能力使用率大于或等于第一阈值、所述终端的计算资源使用率大于或等于第二阈值、所述终端的负载大于或等于第三阈值、所述终端的资源小于或等于第四阈值。

9. 根据权利要求 6 所述的 AI 模型分发方法, 其中,

所述第二 AI 模型为基于至少一个所述终端的训练结果数据对所述第一 AI 模型进行训练得到的; 所述训练结果数据为所述终端对所述第一 AI 模型进行训练得到的。

10. 根据权利要求 6 所述的 AI 模型分发方法, 其中,

5 所述第二 AI 模型的信息包括以下至少一项: 标识、功能描述信息、拓扑结构、参数信息、使用的资源信息。

11. 根据权利要求 5 或 9 所述的 AI 模型分发方法, 其中,

所述训练结果数据包括更新后的 AI 模型的描述信息。

12. 根据权利要求 1 或 2 所述的 AI 模型分发方法, 其中, 所述网络侧设备基于至少一个终端的第一信息, 向各个所述终端发送第一 AI 模型的信息之前, 还包括:

10 所述网络侧设备接收至少一个所述终端的请求消息, 所述请求消息用于请求所述网络侧设备发送所述第一 AI 模型的信息, 所述请求消息包括: 所述终端的第一信息。

13. 根据权利要求 1 或 2 所述的 AI 模型分发方法, 其中, 所述网络侧设备基于至少一个终端的第一信息, 向各个所述终端发送第一 AI 模型的信息之前, 还包括:

15 所述网络侧设备建立与所述终端的第一会话连接; 所述第一会话连接用于传输所述第一 AI 模型的信息。

14. 根据权利要求 5 所述的 AI 模型分发方法, 其中, 所述方法还包括:

所述网络侧设备建立与所述终端的第二会话连接, 所述第二会话连接用于传输所述训练结果数据。

20 15. 一种人工智能 AI 模型接收方法, 包括:

终端接收网络侧设备发送的第一 AI 模型的信息, 所述第一 AI 模型与所述终端的第一信息匹配, 所述第一信息包括以下至少一项: 所述终端的能力信息、AI 模型的需求信息和传输方式;

所述终端基于所述第一 AI 模型进行业务处理。

25 16. 根据权利要求 15 所述的 AI 模型接收方法, 其中, 所述方法还包括:

所述终端基于训练数据对所述第一 AI 模型进行训练, 得到训练结果数据;

所述终端向所述网络侧设备发送所述训练结果数据; 所述训练结果数据用于所述网络侧设备进行 AI 模型训练。

17. 根据权利要求 15 或 16 所述的 AI 模型接收方法, 其中,

30 所述终端的能力信息包括以下至少一项: 计算能力、存储能力、计算能力的使用量、当前剩余计算能力、当前可用计算能力、存储能力的使用量、当前剩余存储能力、当前可用存储能力、可处理的业务信息。

18. 根据权利要求 15 或 16 所述的 AI 模型接收方法, 其中,

所述 AI 模型的需求信息包括以下至少一项: 服务需求信息、AI 模型的描述信息;

35 所述 AI 模型的描述信息包括: 标识、功能描述信息、拓扑结构、参数信息、使用的资

源信息。

19. 根据权利要求 15 或 16 所述的 AI 模型接收方法，其中，所述方法还包括：

所述终端接收所述网络侧设备发送的更新后的第二 AI 模型的信息，所述第二 AI 模型为基于第二信息得到的；所述第二信息包括以下至少一项：各个所述终端的能力更新信息、服务需求更新信息、模型更新信息。

20. 根据权利要求 19 所述的 AI 模型接收方法，其中，

所述能力更新信息包括以下至少一项的更新信息：计算能力、存储能力、计算能力的使用量、当前剩余计算能力、当前可用计算能力、存储能力的使用量、当前剩余存储能力、当前可用存储能力、可处理的业务信息；和/或

所述模型更新信息包括以下至少一项的更新信息：标识、功能描述信息、拓扑结构、参数信息、使用的资源信息。

21. 根据权利要求 19 所述的 AI 模型接收方法，其中，

所述更新后的第二 AI 模型的信息为触发 AI 模型更新的情况下发送的，所述 AI 模型更新为基于第三信息触发的，所述第三信息包括以下至少一项：所述终端的能力更新指示信息、所述网络侧设备更新 AI 模型的指示信息；所述终端的能力更新指示信息用于指示以下至少一项：所述终端的计算能力使用率大于或等于第一阈值、所述终端的计算资源使用率大于或等于第二阈值、所述终端的负载大于或等于第三阈值、所述终端的资源小于或等于第四阈值。

22. 根据权利要求 19 所述的 AI 模型接收方法，其中，

所述第二 AI 模型为基于至少一个所述终端的训练结果数据对所述第一 AI 模型进行训练得到的；所述训练结果数据为所述终端对所述第一 AI 模型进行训练得到的。

23. 根据权利要求 19-22 任一项所述的 AI 模型接收方法，其中，

所述第二 AI 模型的信息包括以下至少一项：标识、功能描述信息、拓扑结构、参数信息、使用的资源信息。

24. 根据权利要求 16 或 22 任一项所述的 AI 模型接收方法，其中，

所述训练结果数据包括更新后的 AI 模型的描述信息。

25. 根据权利要求 15 或 16 所述的 AI 模型接收方法，其中，所述终端接收网络侧设备发送的第一 AI 模型的信息之前，还包括：

所述终端向所述网络侧设备发送请求消息，所述请求消息用于请求所述网络侧设备发送所述第一 AI 模型的信息，所述请求消息包括：所述终端的第一信息。

26. 根据权利要求 15 或 16 所述的 AI 模型接收方法，其中，所述终端接收网络侧设备发送的第一 AI 模型的信息之前，还包括：

所述终端建立与所述网络侧设备的第一会话连接；所述第一会话连接用于传输所述第一 AI 模型的信息。

27. 根据权利要求 16 所述的 AI 模型接收方法，其中，所述方法还包括：

所述终端建立与所述网络侧设备的第二会话连接，所述第二会话连接用于传输所述训练结果数据。

28. 一种 AI 模型分发装置，包括：

5 发送模块，用于在触发 AI 模型分发的情况下，基于至少一个终端的第一信息，向各个所述终端发送第一 AI 模型的信息；所述终端的第一信息与所述终端对应的第一 AI 模型匹配；所述第一信息包括以下至少一项：所述终端的能力信息、AI 模型的需求信息和传输方式。

29. 一种 AI 模型接收装置，包括：

10 接收模块，用于接收网络侧设备发送的第一 AI 模型的信息，所述第一 AI 模型与终端的第一信息匹配，所述第一信息包括以下至少一项：所述终端的能力信息、AI 模型的需求信息和传输方式；

处理模块，用于基于所述第一 AI 模型进行业务处理。

15 30. 一种网络侧设备，包括处理器和存储器，所述存储器存储可在所述处理器上运行的程序或指令，所述程序或指令被所述处理器执行时实现如权利要求 1 至 14 任一项所述的 AI 模型分发方法的步骤。

31. 一种终端，包括处理器和存储器，所述存储器存储可在所述处理器上运行的程序或指令，所述程序或指令被所述处理器执行时实现如权利要求 15 至 27 任一项所述的 AI 模型接收方法的步骤。

20 32. 一种可读存储介质，所述可读存储介质上存储程序或指令，所述程序或指令被处理器执行时实现如权利要求 1 至 14 任一项所述的 AI 模型分发方法，或者实现如权利要求 15 至 27 任一项所述的 AI 模型接收方法的步骤。

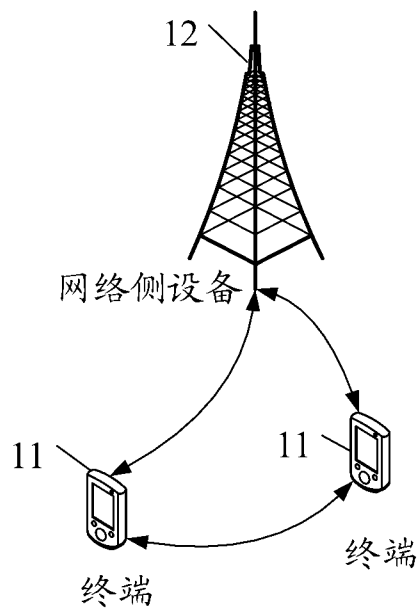


图 1

在触发AI模型分发的情况下，网络侧设备基于至少一个终端的第一信息，向各个终端发送第一AI模型的信息；终端的第一信息与终端对应的第一AI模型匹配；第一信息包括以下至少一项：终端的能力信息、AI模型的需求信息和传输方式

101

图 2

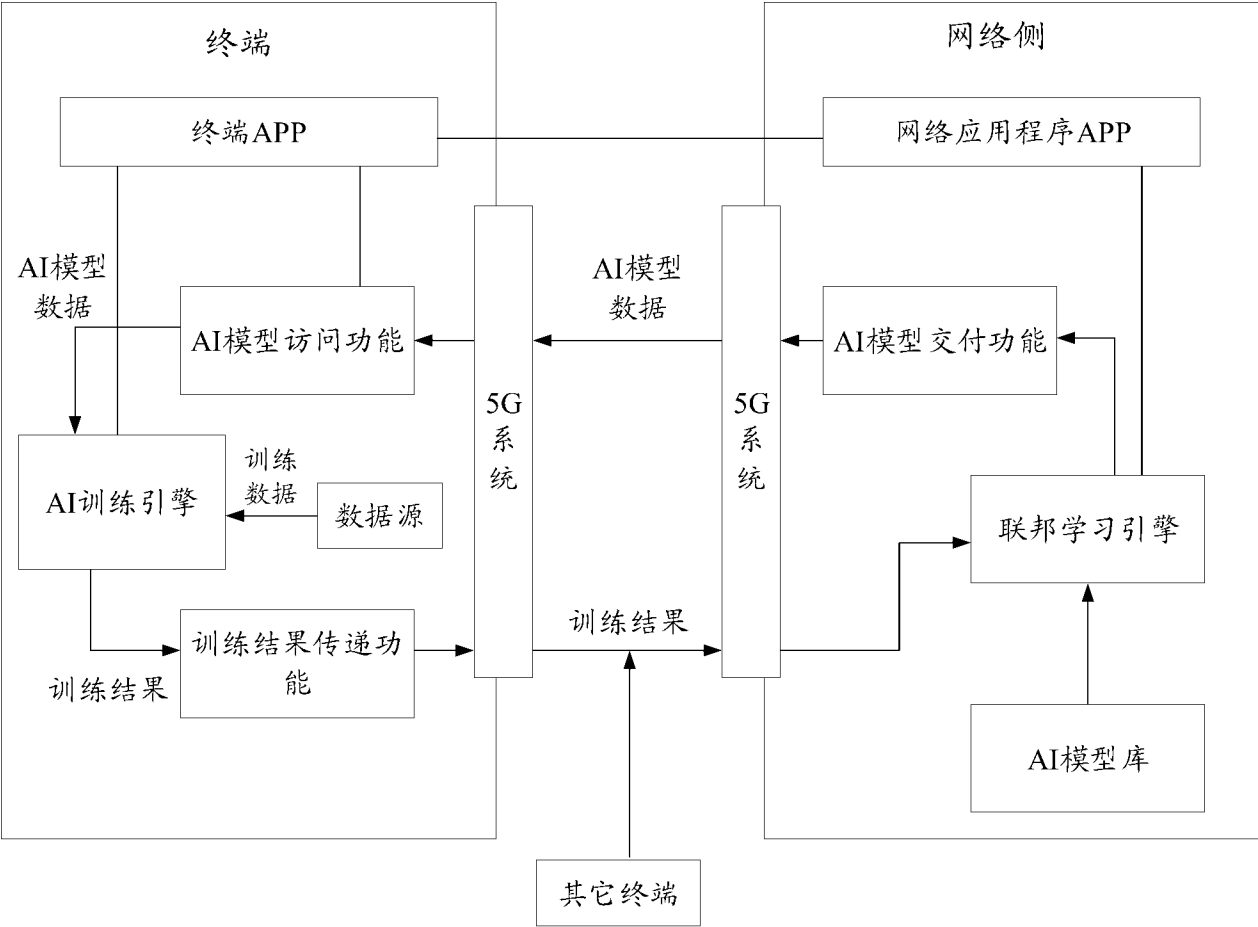


图 3

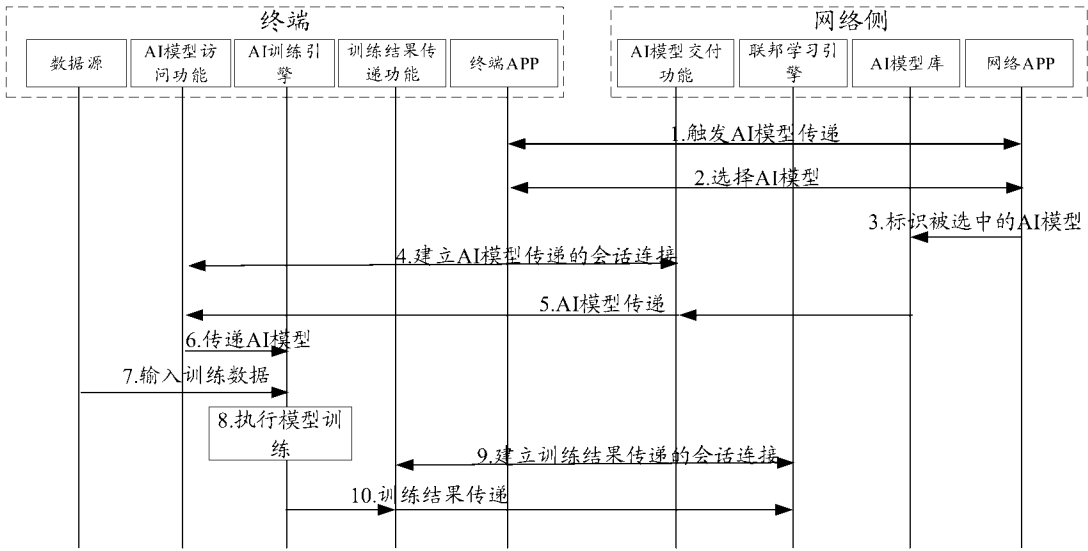


图 4



图 5

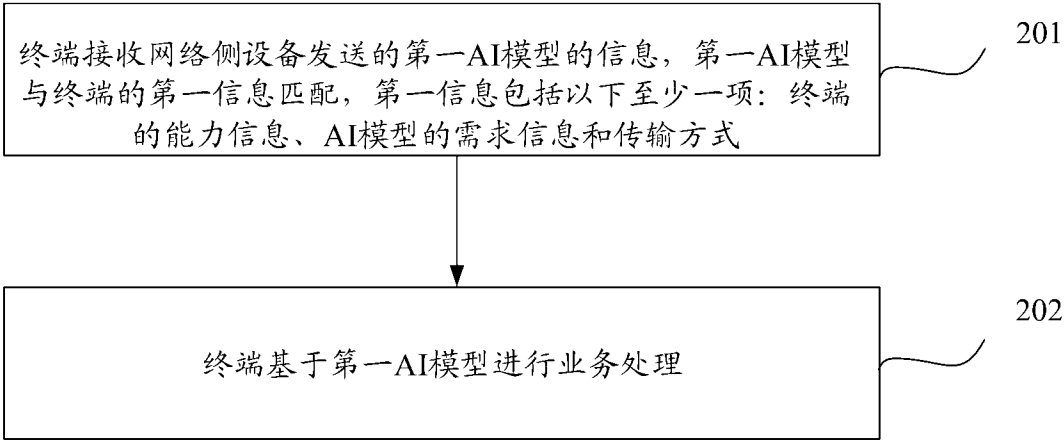


图 6

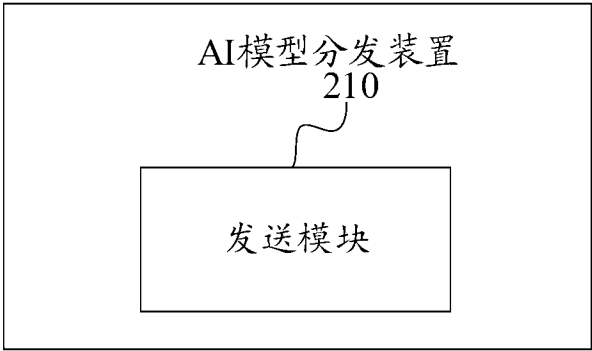


图 7

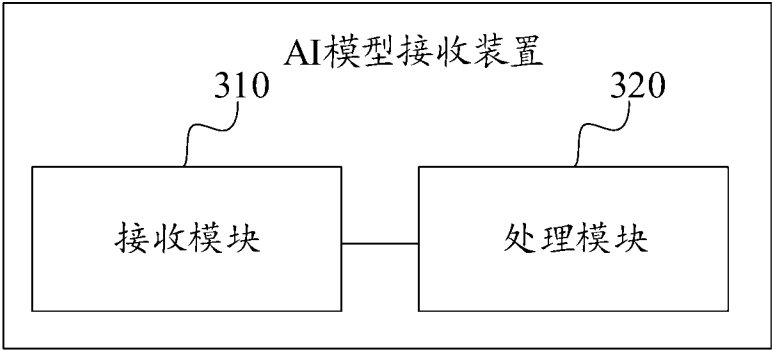


图 8

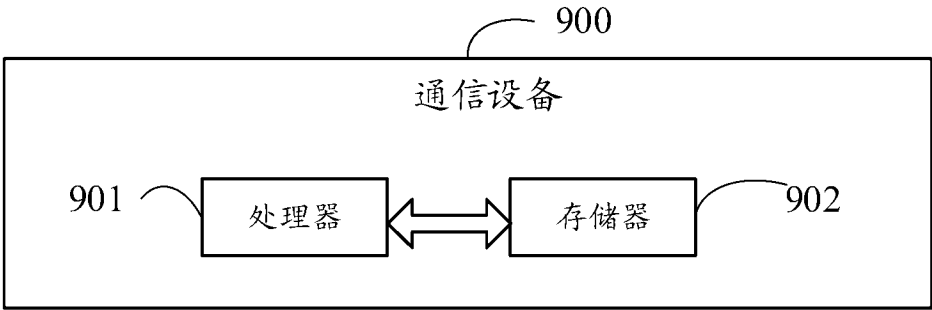


图 9

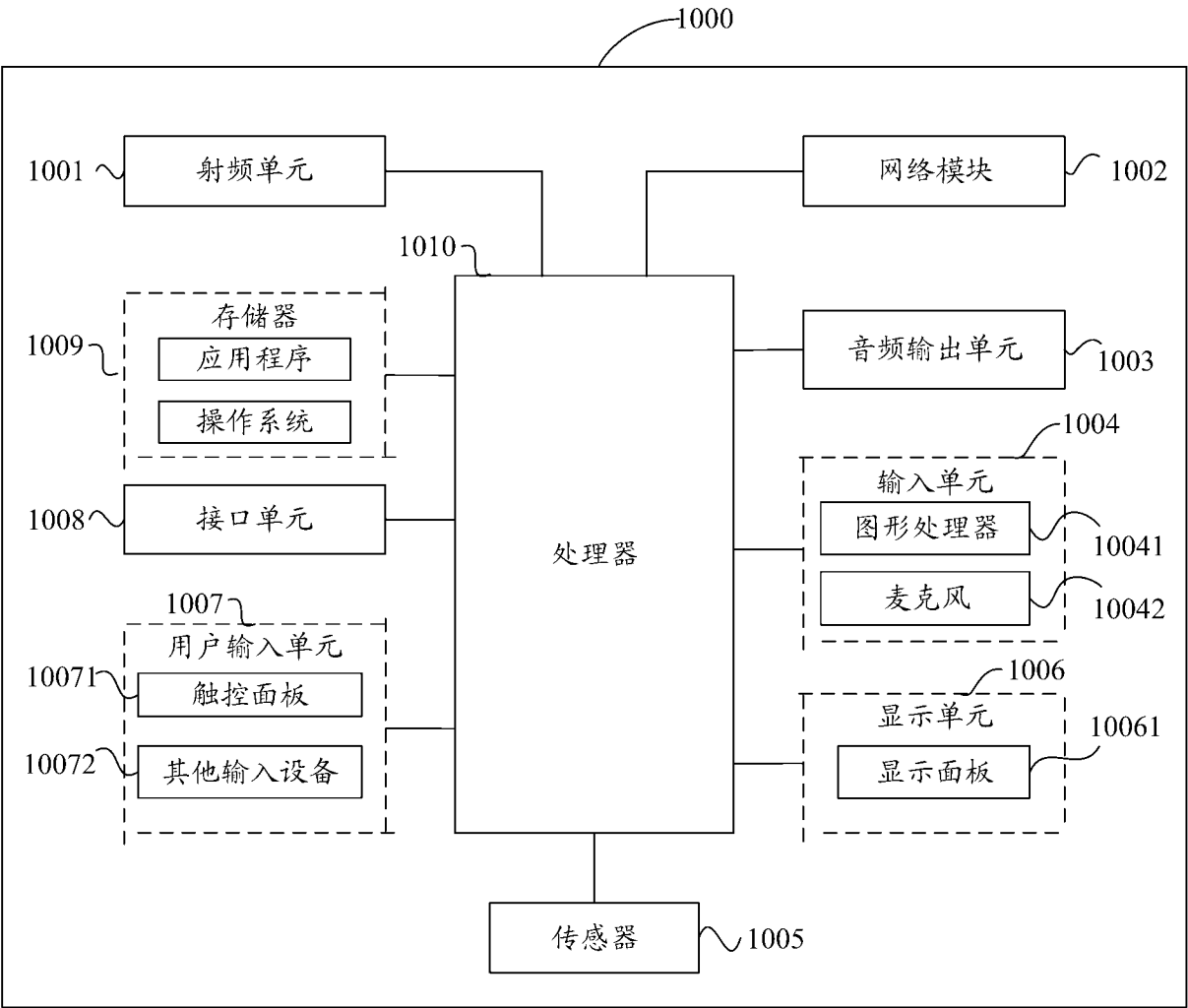


图 10

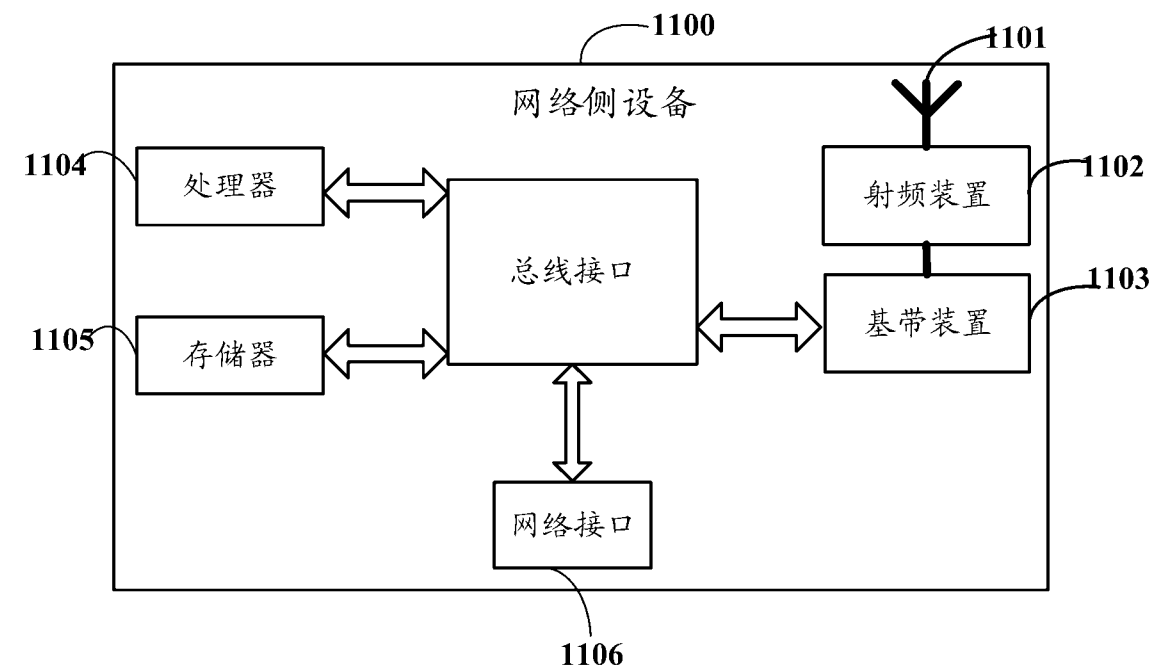


图 11

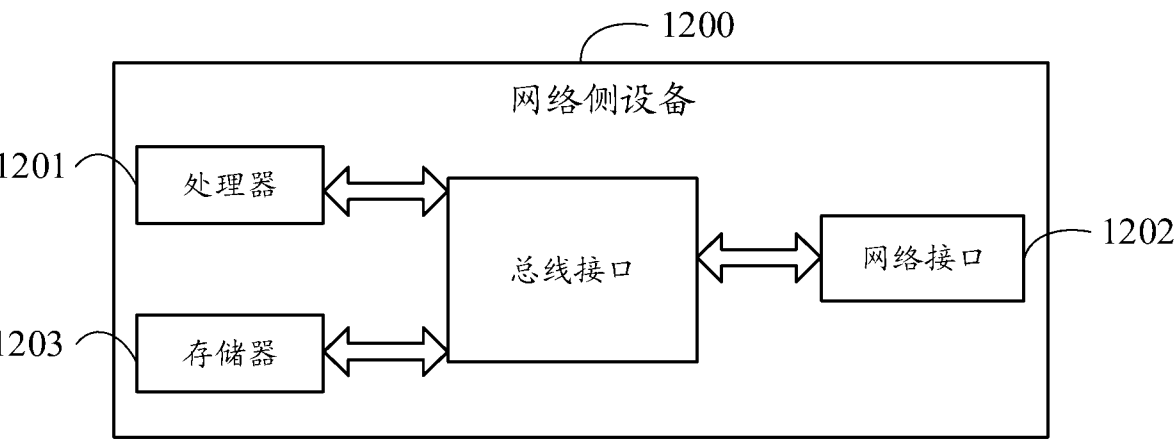


图 12

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2024/071560

A. CLASSIFICATION OF SUBJECT MATTER

H04W 16/22(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)

H04B 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; ENTXTC; ENTXT; DWPI; 3GPP; 机器学习, 人工智能, AI模型, 触发, 请求, 上报, 能力, 分发, 分配, 配置, 配发, 更新, 训练, machine learning, ML, artificial intelligence, AI, models, trigger+, request+, capabilities, distribut+, assign+, configur+, dispens+, updat+, train+

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2021142609 A1 (GUANGDONG OPPO MOBILE TELECOMMUNICATIONS CORP., LTD.) 22 July 2021 (2021-07-22) description, page 4, paragraph 4 to page 5, paragraph 5, and page 8, paragraph 3 to page 11, third-to-last paragraph	1-32
A	CN 114091679 A (HUAWEI TECHNOLOGIES CO., LTD.) 25 February 2022 (2022-02-25) entire document	1-32
A	CN 114745712 A (CHINA MOBILE COMMUNICATION LTD., RESEARCH INSTITUTE; CHINA MOBILE COMMUNICATION GROUP CO., LTD.) 12 July 2022 (2022-07-12) entire document	1-32
A	WO 2022161230 A1 (DATANG MOBILE COMMUNICATIONS EQUIPMENT CO., LTD.) 04 August 2022 (2022-08-04) entire document	1-32



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

01 April 2024

Date of mailing of the international search report

15 April 2024

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
China No. 6, Xitucheng Road, Jimenqiao, Haidian District,
Beijing 100088

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2024/071560

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
WO	2021142609	A1	22 July 2021	EP	4087343	A1	09 November 2022
				EP	4087343	A4	11 January 2023
				US	2022342713	A1	27 October 2022
				CN	114930945	A	19 August 2022
CN	114091679	A	25 February 2022	WO	2022041947	A1	03 March 2022
CN	114745712	A	12 July 2022	None			
WO	2022161230	A1	04 August 2022	EP	4287574	A1	06 December 2023

A. 主题的分类

H04W 16/22(2009.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

H04B H04W

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

CNTEXT;CNKI;ENTXTC;ENTXT;DWPI;3GPP:机器学习, 人工智能, AI模型, 触发, 请求, 上报, 能力, 分发, 分配, 配置, 配发, 更新, 训练, machine learning, ML, artificial intelligence, AI, models, trigger+, request+, capabilities, distribut+, assign+, configur+, dispens+, updat+, train+

C. 相关文件

类型*	引用文件, 必要时, 指明相关段落	相关的权利要求
X	WO 2021142609 A1 (GUANGDONG OPPO MOBILE TELECOMMUNICATIONS CO-RP. LTD.) 2021年7月22日 (2021 - 07 - 22) 说明书第4页第4段-第5页第5段, 第8页第3段-第11页倒数第3段	1-32
A	CN 114091679 A (华为技术有限公司) 2022年2月25日 (2022 - 02 - 25) 全文	1-32
A	CN 114745712 A (中国移动通信有限公司研究院 中国移动通信集团有限公司) 2022年7月12日 (2022 - 07 - 12) 全文	1-32
A	WO 2022161230 A1 (DATANG MOBILE COMMUNICATIONS EQUIPMENT CO., LTD.) 2022年8月4日 (2022 - 08 - 04) 全文	1-32

☐ 其余文件在C栏的续页中列出。

☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“D” 申请人在国际申请中引证的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2024年4月1日

国际检索报告邮寄日期

2024年4月15日

ISA/CN的名称和邮寄地址

中国国家知识产权局
中国北京市海淀区蓟门桥西土城路6号 100088

授权官员

胡绍芹

电话号码 (+86) 010-62412274

PCT/ISA/210 表(第2页) (2022年7月)

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2024/071560

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
WO	2021142609	A1	2021年7月22日	EP	4087343	A1	2022年11月9日
				EP	4087343	A4	2023年1月11日
				US	2022342713	A1	2022年10月27日
				CN	114930945	A	2022年8月19日
CN	114091679	A	2022年2月25日	WO	2022041947	A1	2022年3月3日
CN	114745712	A	2022年7月12日	无			
WO	2022161230	A1	2022年8月4日	EP	4287574	A1	2023年12月6日