

(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
2025 年 5 月 30 日 (30.05.2025)



(10) 国际公布号
WO 2025/108271 A1

(51) 国际专利分类号:
G06N 5/04 (2023.01)

(21) 国际申请号: PCT/CN2024/133006

(22) 国际申请日: 2024 年 11 月 19 日 (19.11.2024)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202311551035.X 2023年11月20日 (20.11.2023) 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 523863 (CN)。

(72) 发明人: 周通(ZHOU, Tong); 中国广东省东莞市长安镇维沃路1号 523863 (CN)。

(74) 代理人: 北京励诚知识产权代理有限公司(BEIJING LISENG IP ATTORNEYS, CO., LTD.); 中国北京市丰台区汽车博物馆东路1号院诺德中心6号楼702 100160 (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。

(54) Title: MODEL REASONING METHOD AND APPARATUS IN COMMUNICATION SYSTEM, AND DEVICE AND MEDIUM

(54) 发明名称: 通信系统中的模型推理方法、装置、设备以及介质

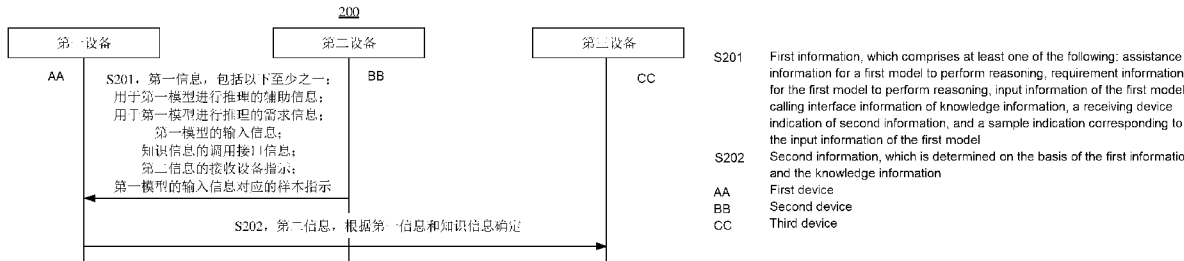


图 2

(57) Abstract: The present application belongs to the field of communications. Disclosed are a model reasoning method and apparatus in a communication system, and a device and a medium. The method in the embodiments of the present application comprises: a first device receiving first information sent by a second device, wherein at least one of knowledge information and calling interface information of the knowledge information is stored in the first device, and the knowledge information comprises at least one of the following: a knowledge graph and a knowledge vector library; and the first device sending second information to a third device, wherein a first model is deployed on the third device, the second information is determined on the basis of the first information and the knowledge information stored in or called by the first device, and the first information comprises at least one of the following: assistance information for the first model to perform reasoning, requirement information for the first model to perform reasoning, input information of the first model, the calling interface information of the knowledge information, a receiving device indication of the second information, and a sample indication corresponding to the input information of the first model.

(57) 摘要: 本申请公开了一种通信系统中的模型推理方法、装置、设备及介质, 属于通信领域, 本申请实施例的方法包括: 第一设备接收第二设备发送的第一信息, 第一设备上存储有知识信息和知识信息的调用接口信息中的至少之一, 其中, 知识信息包括以下至少之一: 知识图谱、知识向量库; 第一设备向第三设备发送第二信息, 第三设备上部署第一模型, 第二信息是根据第一信息和第一设备上存储的或调用的知识信息确定的; 其中, 第一信息包括以下至少之一: 用于第一模型进行推理的辅助信息; 用于第一模型进行推理的需求信息; 第一模型的输入信息; 知识信息的调用接口信息; 第二信息的接收设备指示; 第一模型的输入信息对应的样本指示。

(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))。

通信系统中的模型推理方法、装置、设备以及介质

相关申请的交叉引用

5 本申请要求于 2023 年 11 月 20 日提交中国专利局、申请号为 202311551035.X、发明名称为“通信系统中的模型推理方法、装置、设备以及介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

10 本申请属于通信技术领域，具体涉及一种通信系统中的模型推理方法、装置、设备以及介质。

背景技术

15 在移动通信系统中，开始有越来越多的用例结合人工智能（AI）。例如在物理层有基于 AI 的 CSI（channel state information）预测和反馈压缩，基于 AI 的波束管理，基于 AI 的定位等。为了使不同通信场景具有普适性，提出了将大模型应用到通信网络中来解决通信系统中的问题，但是，大模型通常在解决通用问题上的性能较优，但是通信系统中存在许多不同的问题，例如，波束赋形，资源分配，信道预测等，因此，如何保证大模型在解决具体的下游问题时的性能是一项亟需解决的问题。

发明内容

20 本申请实施例提供一种通信系统中的模型推理方法、装置、设备以及介质，能够提升模型的性能。

第一方面，提供了一种通信系统中的模型推理方法，该方法包括：

25 第一设备接收第二设备发送的第一信息，其中，所述第一设备上存储有知识信息和知识信息的调用接口信息中的至少之一，其中，所述知识信息包括以下至少之一：知识图谱、知识向量库；

所述第一设备向第三设备发送第二信息，所述第三设备上部署第一模型，所述第二信息是根据所述第一信息和所述第一设备上存储的或调用的知识信息确定的；

其中，所述第一信息包括以下至少之一：

30 用于所述第一模型进行推理的辅助信息；

用于所述第一模型进行推理的需求信息；

所述第一模型的输入信息；

知识信息的调用接口信息；

所述第二信息的接收设备指示；

35 所述第一模型的输入信息对应的样本指示。

第二方面，提供了一种通信系统中的模型推理方法，该方法包括：

第二设备向第一设备发送第一信息，其中，所述第一设备上存储有知识信息和知识信息的调用接口信息中的至少之一，其中，所述知识信息包括以下至少之一：知识图谱、知识向量库；

40 其中，所述第一信息包括以下至少之一：

用于第一模型进行推理的辅助信息；

用于第一模型进行推理的需求信息；

第一模型的输入信息；

知识信息的调用接口信息；

所述第二信息的接收设备指示；

所述第一模型的输入信息对应的样本指示。

第三方面，提供了一种通信系统中的模型推理方法，包括：

5 第三设备从第一设备或目标设备获取第二信息，其中，所述第三设备上部署第一模型，所述第一设备或所述目标设备上存储有知识信息和知识信息的调用接口信息中的至少之一，其中，所述目标设备是使用所述调用接口信息所访问的设备，所述知识信息包括以下至少之一：知识图谱、知识向量库；

其中，所述第二信息根据第一信息和所述第一设备上存储或调用的知识信息确定；

其中，所述第一信息包括以下至少之一：

10 用于所述第一模型进行推理的辅助信息；

用于所述第一模型进行推理的需求信息；

所述第一模型的输入信息；

知识信息的调用接口信息；

所述第二信息的接收设备指示；

15 所述第一模型的输入信息对应的样本指示。

第四方面，提供了一种通信装置，包括：

通信单元，用于接收第二设备发送的第一信息，其中，所述通信装置上存储有知识信息和知识信息的调用接口信息中的至少之一，其中，所述知识信息包括以下至少之一：知识图谱、知识向量库；以及

20 向第三设备发送第二信息，所述第三设备上部署第一模型，所述第二信息是根据所述第一信息和所述第一设备上存储的或调用的知识信息确定的；

其中，所述第一信息包括以下至少之一：

用于所述第一模型进行推理的辅助信息；

用于所述第一模型进行推理的需求信息；

25 所述第一模型的输入信息；

知识信息的调用接口信息；

所述第二信息的接收设备指示；

所述第一模型的输入信息对应的样本指示。

第五方面，提供了一种通信装置，包括：

30 通信单元，用于向第一设备发送第一信息，其中，所述第一设备上存储有知识信息和知识信息的调用接口信息中的至少之一，其中，所述知识信息包括以下至少之一：知识图谱、知识向量库；

其中，所述第一信息包括以下至少之一：

用于第一模型进行推理的辅助信息；

35 用于第一模型进行推理的需求信息；

第一模型的输入信息；

知识信息的调用接口信息；

所述第二信息的接收设备指示；

所述第一模型的输入信息对应的样本指示。

40 第六方面，提供了一种通信装置，包括：

通信单元，用于从第一设备或目标设备获取第二信息，其中，所述第三设备上部署第一模型，所述第一设备或所述目标设备上存储有知识信息和知识信息的调用接口信息中的至少之一，其中，所述目标设备是使用所述调用接口信息所访问的设备，所述知识信息包括以下至少之一：知识图谱、知识向量库；

45 其中，所述第二信息根据第一信息和所述第一设备上存储或调用的知识信息确定；

其中，所述第一信息包括以下至少之一：

用于所述第一模型进行推理的辅助信息；
用于所述第一模型进行推理的需求信息；
所述第一模型的输入信息；
知识信息的调用接口信息；

5 所述第二信息的接收设备指示；
所述第一模型的输入信息对应的样本指示。

第七方面，提供了一种通信设备，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至第三方面中的任一方面所述的方法的步骤。

10 第八方面，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至第三方面中的任一方面所述的方法的步骤。

15 第九方面，提供了一种无线通信系统，包括：第一设备、第二设备及第三设备，所述第一设备可用于执行如第一方面所述的方法的步骤，所述第二设备可用于执行如第二方面所述的方法的步骤，所述第三设备可用于执行如第三方面所述的方法的步骤。

第十方面，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，实现如第一方面至第三方面中的任一方面中所述的方法。

20 第十一方面，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，所述程序/程序产品被至少一个处理器执行以实现如第一方面至第三方面中的任一方面所述的方法。

在本申请实施例中，第二设备可以通过向第一设备发送第一信息，辅助第三设备获得第二信息，进一步地，第三设备可以使用第二信息辅助进行模型的推理，有利于提升模型的推理性能。

25

附图说明

图1是本申请实施例提供的一种通信系统的示意性图。

图2是本申请实施例提供的一种通信系统中的模型推理方法的示意性图。

图3是本申请实施例提供的一种隐藏变量的示意性图。

30 图4至图14是本申请实施例提供的模型推理方法的示意性交互图。

图15是本申请实施例提供的一种通信装置的示意性图。

图16是本申请实施例提供的另一种通信装置的示意性图。

图17是本申请实施例提供的又一种通信装置的示意性图。

图18是本申请实施例提供的一种通信设备的示意性图。

35 图19是本申请实施例提供的一种终端的硬件结构图。

图20是本申请实施例提供的一种网络侧设备的硬件结构图。

图21是本申请实施例提供的另一种网络侧设备的硬件结构图。

具体实施方式

40 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。

本申请的术语“第一”、“第二”等是用于区别类似的对象，而不适用于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
45 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，本申请中的“或”表示所连接对象的至少其中之一。例如“A 或 B”涵盖三种方案，即，方案一：包括 A 且不包括 B；方案二：包括 B 且不包括 A；方案三：既包括 A 又包括 B。字符“/”一般表示前后关联对象是一种“或”的关系。

5 本申请的术语“指示”既可以是一个直接的指示（或者说显式的指示），也可以是一个间接的指示（或者说隐含的指示）。其中，直接的指示可以理解为，发送方在发送的指示中明确告知了接收方具体的信息、需要执行的操作或请求结果等内容；间接的指示可以理解为，接收方根据发送方发送的指示确定对应的信息，或者进行判断并根据判断结果确定需要执行的操作或请求结果等。

10 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)或其他系统。本申请实
15 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和无线电技术，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 系统以外的系统，如第 6 代(6th Generation, 6G)通信系统。

20 图 1 示出本申请实施例可应用的一种无线通信系统的框图。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)、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)、飞行器(flight vehicle)、车载设备(Vehicle User Equipment, VUE)、船载设备、行人终端(Pedestrian User Equipment, 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。其中，车载设备也可以称为车载终端、车载控制器、车载模块、车载部件、车载芯片或车载单元等。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。
25 30

35 终端也可以称为用户设备(User Equipment, UE)、终端设备、接入终端、用户单元、用户站、移动站、移动台、远方站、远程终端、移动设备、用户终端、终端、无线通信设备、用户代理或用户装置等。

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(Radio Access Network, RAN)设备、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、无线局域网(Wireless Local Area Network, WLAN)接入点(Access Point, AS)或无线保真(Wireless Fidelity, WiFi)节点等。其中，基站可被称为节点 B(Node B, NB)、演进节点 B(Evolved Node B, eNB)、下一代节点 B(the next generation Node B, gNB)、新空口节点 B(New Radio Node B, NR Node B)、接入点、中继站(Relay Base Station, RBS)、服务基站(Serving Base Station, SBS)、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(home Node B, HNB)、家用演进型 B 节点(home evolved Node B)、发送接收点(Transmission Reception
40 45

Point, TRP) 或所属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)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。但包括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)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。

为便于理解理解本申请实施例, 对本申请相关的技术进行说明。

一、知识图谱

知识图谱是一种利用图结构或拓扑来表示和整合数据的知识库, 可以存储实体(对象、事件、情况或抽象概念)之间的关联描述, 同时也可以编码实体之间的语义关系。知识图谱的应用前景非常广阔, 既能提高信息检索、搜索引擎、推荐系统等数据服务的效果, 还支持自然语言问答、对话、推理等智能交互, 具备表示能力强、灵活性高、可计算性强和跨领域性强等特点。

知识图谱以明确和结构化的方式存储了大量的知识, 可以用于增强大模型的知识意识。在推理阶段将知识图谱纳入大模型中, 通过从知识图谱检索知识, 可以显著提高大模型在访问领域特定知识方面的性能。

在通信领域, 主要利用网络数据知识图谱进行知识表示、关联关系分析和深度挖掘, 为通信系统的智能化提供有效的知识规则与知识计算支撑。

通信系统的网络结构、终端类型、终端行为、数据业务需求、系统资源都具有高动态性、时效性强及相互耦合性等特点, 移动通信数据面临诸多挑战, 数据分散获取困难, 种类繁多结构复杂, 关联复杂挖掘困难等等。

利用知识图谱可以有效厘清数据字段与通信网络指标之间的各种关系, 并在关系建立的基础上进一步深度挖掘, 比如量化关系的关联程度、表征数据字段与指标的特征属性等等。

在一些实现方式中，知识图谱可以通过实体之间的相关性表示，一种表示方式可以为：

（实体 1，实体 1 和实体 2 之间的相关性系数，实体 2）。

例如，实体 1 是小区吞吐量，实体 2 是最强波束的 L1-RSRP，相关性系数为 0.5。

5 在另一些实现方式中，知识图谱可以通过主体属性对象（Subject, Predicate, Object, SPO）三元组表示，例如，（小区 id，赋形码本，码本指示）的 SPO 三元组。

二、知识向量库

知识向量库是用来存储、检索、分析向量的数据库。之所以称之为数据库，是因为它有下面几个特征：

10 a) 提供标准的访问接口，降低用户的使用门槛；

b) 提供高效的数据组织、检索和分析的能力。一般用户在存储和检索向量的同时，还需要管理结构化的数据，如支持传统数据库对结构化数据的管理能力。

因此，知识向量库可以简单理解为用于存储模型输入特征向量的数据库。

15 例如，在利用图片搜索图片，或者，利用语音搜索语音的时候，在知识向量库中存储和对比的并不是图片和语音片段，而是通过深度学习等算法提取出来的“特征”，例如 256 或 512 个浮点数（float）数组，可以用数学中的向量来表示。

在一些情况中，知识向量库可以是一个模型，输入的是一个图片，文字。例如，在通信领域，模型的输入可以是无线信号的测量结果等，输出是一个用于表征该测量结果的特征向量。

20 在本地可以采用如下方法使用知识图谱或知识向量库：

1) 直接作为模型的输入；

2) 用知识图谱对模型输入进行处理后，再把处理后的数据和原始的模型输入一起输入到模型中。

25 3) 用知识向量库对模型输入进行处理后，再把处理后的数据和原始的模型输入一起输入到模型中，或者直接使用处理后的数据输入到模型中。

三、提示工程

30 提示工程是自然语言处理（Natural Language Processing, NLP）中的一种技术，可以制作文本片段、提示或模板，用于指导预训练大语言模型产生特定任务或应用程序的高质量输出，广泛用于问答、摘要、翻译、情感分析和文本生成等领域。提示工程可以利用预训练大语言模型的强大能力来实现各种复杂的自然语言处理任务，减少对标记数据和模型微调的依赖性，降低开发成本和时间，以及提高预训练大语言模型的可解释性和可控性，增加用户信任和满意度。常见的提示可以包括如下零样本提示（Zero-Shot Prompting）、少样本提示（Few-Shot Prompting）、思维链提示（Chain of Thought Prompting）等。

35 在提示工程中，任务的描述会被嵌入到输入中，从而引导生成式人工智能解决方案生成所需输出。

40 在移动通信系统中，开始有越来越多的用例结合人工智能（Artificial Intelligence, AI）。例如在物理层有基于 AI 的信道状态信息（Channel State Information, CSI）预测和反馈压缩，基于 AI 的波束管理，基于 AI 的定位等。在一些场景中，也考虑引入基于 AI 的节能，基于 AI 的负载均衡等。未来在移动通信系统中会出现更多的结合 AI 的用例。

近年来，大语言模型在聊天、图像生成等领域获得了广泛关注。因此，考虑将大模型应用到通信网络中，作为一种工具来优化通信网络的性能。但是，通信大模型有两个问题。

45 问题 1：大模型的尺寸很大，如何将大模型部署到例如基站或者终端等存储资源有限的设备进行推理，较为典型的方法是通过量化、剪枝等压缩方法对大模型进行处理后，再部署到通信设备上。

问题 2：大模型往往是解决通用问题，而通信中有很多不同的问题，例如波束赋形，资源分配，信道预测。如何将大模型匹配到具体的下游问题上，目前在通信系统中还没有成熟的方案。

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通信系统中基于人工智能 AI 模型的推理方法进行详细地说明。

图 2 示出了根据本申请实施例的通信系统中的模型推理方法的示意性图。如图 2 所示，该方法 200 包括：

S201，第一设备接收第二设备发送的第一信息，其中，所述第一设备上存储有知识信息和知识信息的调用接口信息中的至少之一；

S202，所述第一设备向第三设备发送第二信息，所述第三设备上部署第一模型，所述第二信息是根据所述第一信息和知识信息确定的。

在一些实施例中，所述第一模型是压缩后的 AI 模型，例如，第一模型是对大模型进行压缩后的 AI 模型，这里的压缩例如可以包括但不限于量化、剪枝（例如，删除模型中的部分层，或者删除某层中的部分节点）等。

需要说明的是，在本申请实施例中，AI 模型也可称为 AI 单元、ML(machine learning)模型、ML 单元、AI 结构、AI 功能、AI 特性、机器学习模型、神经网络、神经网络函数、神经网络功能等，或者，AI 模型也可以是指能够实现与 AI 相关的特定的算法、公式、处理流程、能力等的处理单元，或者 AI 模型可以是针对特定数据集的处理方法、算法、功能、模块或单元，或者 AI 模型可以是运行在图形处理单元（Graphics Processing Unit, GPU）、神经网络处理单元(Neural Processing Unit, NPU)、张量处理单元(Tensor Processing Unit, TPU)、专门应用集成电路（Application Specific Integrated Circuit, ASIC）等 AI/ML 相关硬件上的处理方法、算法、功能、模块或单元，本申请对此不做具体限定。

在本申请实施例中，第一设备可以认为是存储知识相关信息的设备，或称知识存储设备，例如第一设备上存储有知识信息、知识信息的调用接口信息等。第二设备可以认为是需求发起设备，如发起推理需求的设备。第三设备可以认为是推理设备，或者说，AI 模型部署设备。其中，第二设备和第三设备可以是同一设备，或者，也可以是不同设备。

在一些实施例中，知识信息包括但不限于知识图谱和知识向量库中的至少之一。

在一些实施例中，所述第三设备可以利用第二信息扩充第一模型的输入，辅助第一模型进行精准的推理。

因此，在本申请实施例中，通过向存储有知识存储设备获取第二信息，进一步使用模型进行推理时，可以使用第二信息和第一模型的输入信息，从而能够使得模型能够获得更多的输入信息，提升模型的推理性能。

在一些实施例中，所述第一设备为接入网设备（例如基站）或核心网功能，例如具有数据存储功能的核心网功能，例如数据库、数据功能或网络存储功能（Network Repository Function, NRF）或 UDM，或 AI 模型库，或 AI 模型管理功能或终端的第三方服务器等，第二设备是终端，第三设备是终端。

在另一些实施例中，所述第一设备为核心网功能，例如具有数据存储功能的核心网功能，例如数据库、数据功能或网络存储功能（Network Repository Function, NRF）或 UDM，或 AI 模型库，或 AI 模型管理功能或终端的第三方服务器等，第二设备是终端或接入网设备（例如基站），第三设备是接入网设备（例如基站）。

在又一些实施例中，所述第一设备为核心网功能，例如具有数据存储功能的核心网功能，例如数据库、数据功能或网络存储功能（Network Repository Function, NRF）或 UDM，或 AI 模型库，或 AI 模型管理功能或终端的第三方服务器等，第二设备是终端或接入网设备（例如基站）、第三方服务器、核心网功能（例如 AI 控制功能、任务控制功能、协同控制功能或其他核心网功能等），第三设备是核心网的推理功能。

在本申请实施例中，该 AI 控制节点可以用于 AI 相关功能的控制，任务控制节点可以用于服务相关功能的控制，或者任务相关的控制功能，例如，该任务可以包括但不限于网络新能力涉及到多节点场景下连接、计算、数据和算法资源的协同和调配，用于共同完成某个特定的目标，协同控制节点可以用于多个功能（例如通信功能，数据功能，算力功能，算法功能，模型功能）之间的协同管理。

在本申请实施例中，终端和接入网设备之间的信息交互可以通过以下至少一种信令：层 1 信令、层 2 信令、层 3 信令、数据面信令、AI 层信令。终端和核心网设备之间的信息交互可以通过以下至少一种信令中：NAS 信令、数据面信令、AI 层信令。终端和核心网设备之间可以直接进行信息的交互，或者，也可以通过其他设备进行信息的转发，例如通过通信控制功能（例如接入移动管理功能 AMF）、AI 控制功能、任务控制功能、协同控制功能、网络开放功能等进行信息的转发。

应理解，接入网设备和核心网设备，以及核心网设备之间可以直接进行信息的交互，或者，也可以通过其他设备进行信息的转发，例如通过通信控制功能（例如接入移动管理功能 AMF）、AI 控制功能、任务控制功能、协同控制功能、网络开放功能等进行信息的转发。

在一些实施例中，若所述第一设备为接入网设备，所述第二设备为终端，第一信息携带在以下至少一种信令中：层 1 信令、层 2 信令、层 3 信令、数据面信令、AI 层信令。示例性地，该层 1 信令可以包括但不限于 PDCCH。该层 2 信令例如可以包括但不限于下行 MAC CE，该层 3 信令例如可以包括但不限于 RRC 信令。

在另一些实施例中，若所述第一设备为核心网功能，所述第二设备为终端，第一信息携带在以下至少一种信令中：NAS 信令、数据面信令、AI 层信令。

在一些实施例中，若所述第一设备为接入网设备，所述第三设备为终端，第二信息携带在以下至少一种信令中：层 1 信令、层 2 信令、层 3 信令、数据面信令、AI 层信令。示例性地，该层 1 信令可以包括但不限于 PDCCH。该层 2 信令例如可以包括但不限于下行 MAC CE，该层 3 信令例如可以包括但不限于 RRC 信令。

在另一些实施例中，若所述第一设备为核心网功能，所述第三设备为终端，第二信息携带在以下至少一种信令中：NAS 信令、数据面信令、AI 层信令。

在本申请一些实施例中，所述第一设备也可以将第一信息发送给所述第三设备，用于辅助第三设备进行模型推理。例如第一信息可以携带在第二信息中，如第二信息包括第一信息中的内容。

在一些实施例中，所述第一信息包括以下至少之一：

用于所述第一模型进行推理的辅助信息；

用于所述第一模型进行推理的需求信息；

所述第一模型的输入信息；

知识信息的调用接口信息；

所述第二信息的接收设备指示；

所述第一模型的输入信息对应的样本指示。

因此，需求发起设备通过向知识存储设备指示用于模型推理的相关信息（如第一信息），知识存储设备可以基于该第一信息和知识信息确定第二信息，例如，根据第一信息获取专业领域的知识信息，从而得到第二信息，进一步向推理设备发送第二信息，这样推理设备在使用第二信息辅助进行模型的推理，能够实现专业领域任务的精准推理，降低模型推理中的幻觉问题，提升模型的推理性能。

在一些实施例中，接收设备指示可以是第二设备的接收设备的标识信息，用于第一设备确定将第二信息发送给哪个设备。

在一些实施例中，样本指示可以是第一信息中的第一模型的输入信息对应哪个样本，

避免推理设备使用对应不同的样本的第一信息和第二信息进行模型推理，影响推理结果。例如，该样本指示可以是样本标识。

在一些实施例中，所述用于所述第一模型进行推理的辅助信息包括以下至少之一：

5 AI模型的标识信息、AI模型的功能指示、AI模型的特性指示、目标知识图谱指示、目标知识向量库指示、小区标识、带宽部分（Band Width Part, BWP）标识。

在一些实施例中，所述AI模型的标识信息例如可以是AI模型的模型标识(model ID)可以用于AI结构标识、AI算法标识，或者AI模型关联的特定数据集的标识，或者AI/ML相关的特定场景、环境、信道特征、设备的标识，或者AI/ML相关的功能、特性、能力或模块的标识，本申请对此不做具体限定。

10 在一些实施例中，AI模型的特性指示可以显式或隐式指示AI模型支持的特性，例如支持CSI预测和压缩反馈、波束预测、定位、负载均衡、资源分配等。通过向第一设备指示AI模型的特性，可以辅助第一设备获取该特性相关的知识信息，然后根据该知识信息获取相应的第二信息，用于辅助第三设备进行模型的推理。

在本申请实施例中，可以认为在具体配置下的AI模型的特性是AI模型的功能。

15 例如，AI模型的特性为波束预测，AI模型的功能为基站配置为32发送波束下的时间域波束预测。

在一些实施例中，AI模型的功能指示可以显式或隐式指示AI模型支持的功能，例如支持特定配置下的CSI预测和压缩反馈、波束预测、定位、负载均衡、资源分配等。通过向第一设备指示AI模型的功能，可以辅助第一设备获取该功能相关的知识信息，然后

20 根据该知识信息获取相应的第二信息，用于辅助第三设备进行模型的推理。
在一些实施例中，所述目标知识图谱指示可以用于指示第二设备期待使用的知识图谱。通过向第一设备指示期望的目标知识图谱，可以辅助第一设备使用该目标知识图谱获取相应的第二信息，用于辅助第三设备进行模型的推理。

在一些实现方式中，目标知识图谱指示可以显式指示目标知识图谱。

25 示例性地，所述目标知识图谱指示用于以下至少之一：

目标知识图谱中的实体之间的关系；

目标知识图谱中的实体的属性信息；

目标知识图谱的数据结构；

目标知识图谱中的实体之间的相关性；

30 目标知识图谱中的语句的主体属性对象（Subject, Predicate, Object, SPO）三元组；
目标知识图谱文件。

在一些实施例中，目标知识图谱中的实体之间的关系可以采用如下格式表示：

（实体1，关系，实体2）。

可选地，实体之间的关系可以是包括但不限于上下位关系，继承关系等。

35 在一些实施例中，目标知识图谱中的实体的属性信息可以采用如下格式表示：

（实体，属性名，属性值）。

在一些实施例中，可以通过三元组（triple）来表示节点之间的关系，例如一种关系可以是（节点1，边，节点2）。

在一些实施例中，目标知识图谱中的实体之间的相关性可以采用如下格式表示：

40 （实体1，实体1和实体2之间的相关性系数，实体2）。

例如，实体1是小区吞吐量，实体2是最强波束的L1-RSRP，相关性系数为0.8。

在一些实施例中，目标知识图谱中的语句的SPO三元组用于描述语句的主语、谓语和宾语。

在另一些实现方式中，目标知识图谱指示可以隐式指示目标知识图谱。

45 示例性地，所述目标知识图谱指示包括但不限于以下至少之一：

知识图谱标识；

知识图谱文件地址。

在一些实施例中，可以预定义或预配置（例如网络侧设备预先指示）多个知识图谱标识，每个知识图谱标识对应一个知识图谱，例如，第二设备可以在该多个知识图谱标识中选择一个或多个知识图谱标识指示给第一设备，进一步地，该第一设备可以根据指示的知识图谱标识确定对应的知识图谱，进一步基于该知识图谱确定第二信息。

在一些实施例中，可以预定义或预配置（例如网络侧设备预先指示）多个知识图谱文件地址，每个知识图谱文件地址对应一个知识图谱文件，例如，第二设备可以在该多个知识图谱文件地址中选择一个或多个知识图谱文件地址指示给第一设备，进一步地，该第一设备可以根据指示的知识图谱文件地址获取对应的知识图谱文件，进一步基于该知识图谱文件确定第二信息。

在一些实施例中，以 AI 模型的任务是波束预测为例，知识图谱的使用方式可以为：根据小区标识和目标任务，利用知识图谱可以检索出小区发送波束配置，例如发送波束的指向或赋形码本。

在一些实现方式中，目标知识向量库指示可以隐式指示第二设备期望使用的目标知识向量库。通过向第一设备指示期望的目标知识向量库，可以辅助第一设备使用该目标知识向量库获取相应的第二信息，用于辅助第三设备进行模型的推理。

示例性地，所述目标知识向量库指示包括但不限于以下至少之一：

知识向量库标识；

知识向量库文件地址。

在一些实施例中，可以预定义或预配置（例如网络侧设备预先指示）多个知识向量库标识，每个知识向量库标识对应一个知识向量库，例如，第二设备可以在该多个知识向量库标识中选择一个或多个知识向量库标识指示给第一设备，进一步地，该第一设备可以根据指示的知识向量库标识确定对应的知识向量库，进一步基于该知识向量库确定第二信息。

在一些实施例中，可以预定义或预配置（例如网络侧设备预先指示）多个知识向量库文件地址，每个知识向量库文件地址对应一个知识向量库文件，例如，第二设备可以在该多个知识向量库文件地址中选择一个或多个知识向量库文件地址指示给第一设备，进一步地，该第一设备可以根据指示的知识向量库文件地址获取对应的知识向量库文件，进一步基于该知识向量库文件确定第二信息。

在一些实施例中，以 AI 模型的任务是波束预测为例，知识向量库的使用方法可以为：根据小区标识和目标任务，利用知识向量库获得小区发送波束配置，例如发送波束的指向或赋形码本相关的特征向量。

在一些实施例中，该小区标识可以用于第一设备获取该小区标识对应的小区的知识信息，例如小区配置等，进一步地，可以根据该知识信息确定第二信息。

在一些实施例中，该 BWP 标识可以用于第一设备获取该 BWP 标识对应的 BWP 的知识信息，例如 BWP 配置等，进一步地，可以根据该知识信息确定第二信息。

在一些实施例中，所述用于所述第一模型进行推理的需求信息包括以下至少之一：

AI 模型所需适配的目标任务指示；

AI 模型的业务体验质量（Quality of Experience, QoE）要求；

AI 模型的业务处理时延；

AI 模型的业务处理准确度；

AI 模型的业务计算量；

AI 模型的数据处理规模。

例如，第一设备可以根据该目标任务指示选择与该目标任务相关的知识信息，然后根据该知识信息确定第二信息。

又例如，第二设备可以根据 QoE 要求、或处理时延、处理准确度或处理规模等，选

择满足 QoE 要求、或处理时延、处理准确度或处理规模的知识信息，然后根据该知识信息确定第二信息。

在一些实施例中，所述第一模型的输入信息与第一模型所需适配的目标任务有关，以目标任务是波束预测为例，该输入信息可以包括小区标识和小区发送波束的测量质量。

5 在一些实施例中，所述知识信息的调用接口信息包括但不限于如下至少之一：

调用接口名称、调用接口的输入数据格式、调用接口的输出数据格式、调用接口的输入数据。其中，该输入数据格式、输出数据格式、输入数据与领域、任务、场景等相关。

10 在一些实施例中，调用接口可以包括但不限于应用程序接口（Application Programming Interface, API）等访问外部设备的接口。

通信系统会不断地迭代更新，通信系统的知识也会更新。在网络部署之后，更新的系统知识，例如来自 3GPP 等组织的最新知识可以存储在外部的数据库中，此情况下，通过调用 API 接口的方式，为通信系统提供服务，提升了本申请技术方案的适用性。

15 在一些实施例中，第二设备和第三设备都为终端，第一信息可以是在特定情况下发送的，例如：

在发生以下事件中的至少之一时，第二设备向第一设备发送所述第一信息：

事件 1：接收到第一指示，所述第一指示用于指示终端进行模型的激活或切换；

事件 2：所述终端确定进行模型的激活或切换；

事件 3：所述终端确定当前使用的模型不满足需求；

20 事件 4：接收到第二指示，所述第二指示用于指示当前终端使用的模型不满足需求。

在一些实施例中，触发第二设备向第一设备发送第一信息的事件可以是预定义的，或者，网络侧设备配置的，当满足触发第一信息的事件时，第二设备可以将该事件的标识信息发送给网络侧设备，例如，携带在第一信息中发送给网络侧设备，从而网络侧设备可以获知第二设备上发生的事件。

25 在一些实施例中，第一指示可以是网络侧设备发送的，例如，该网络侧设备可以控制 AI 模型的激活或切换，比如，网络侧设备可以在当前使用的模型不满足需求时，向终端发送第一指示，从而第二设备可以根据第一指示发送第一信息。

在一些实施例中，终端可以在当前使用的模型的性能不满足需求时，自主确定进行模型的激活或切换。

30 在一些实施例中，第二指示可以是网络侧设备发送的，例如，网络侧设备可以在当前使用的模型不满足需求时，向终端发送第二指示，用于指示终端当前使用的模型不满足需求，进一步地，该终端可以根据该第二指示自主判断是否进行模型的激活或切换。

35 在一些实施例中，模型不满足需求可以指模型不满足需求信息，例如可以包括但不限于以下至少之一：模型的推理不满足业务体验质量（Quality of Experience, QoE）不满足 QoE 要求，模型的业务处理时延不满足时延需求，模型的业务处理准确度不满足准确度需求。

在本申请一些实施例中，所述第二信息包括以下至少之一：

隐藏变量，所述隐藏变量是对所述第一模型的输入信息进行映射后的特征向量；

提示信息；

40 网络侧设备的配置信息；

场景类型信息；

所述第一模型的输入信息对应的样本指示，或者，确定所述第二信息所使用的第一信息中的样本指示。

45 在一些实施例中，样本指示可以用于指示第二信息是根据哪个样本确定的，该样本指示可以是样本标识。这样，在推理设备获知第一信息和第二信息时，可以通过该样本指示确定第一信息和第二信息对应的样本，进一步在进行模型推理时，可以使用对应相

同样本的第一信息和第二信息，避免使用对应不同样本的第一信息和第二信息进行模型推理，影响推理结果。

在一些实施例中，隐藏变量可以指将具有物理含义的输入信息，经过编码器，映射得到的低维特征，或者，原始输入经过去除冗余后的特征向量。

5 例如，如图 3 所示，自编码器包括编码器和解码器，编码器将输入的图像映射到低维特征（图 3 中的隐藏变量），解码器基于隐藏变量去还原输入图像。其中隐藏变量包含了输入图像的特征信息。

在一些实施例中，提示信息可以用于提示或引导第一模型推理出符合预期的输出，提升模型的推理性能。

10 示例性地，所述提示信息包括以下至少之一：

与所述第一模型适配的目标任务相关的思维链；

与所述目标任务相关的提示文字；

与所述目标任务相关的提示文件；

与所述目标任务相关的提示代码；

15 与所述目标任务相关的思维链标识，用于标识一个思维链；

与所述目标任务相关的提示文字标识，用于标识一段提示文字；

与所述目标任务相关的提示文件标识，用于标识一个提示文件；

与所述目标任务相关的提示代码标识，用于标识一段提示代码；

与所述目标任务相关的提示文件地址。

20 在一些实施例中，所述网络侧设备的配置信息包括但不限于以下至少之一：

网络侧设备的发送天线个数；

网络侧设备的发送波束个数；

网络侧设备的天线端口个数；

网络侧设备的发送功率；

25 网络侧设备的天线增益；

网络侧设备的波束 3dB 带宽；

网络侧设备的间距；

网络侧设备的频率；

系统带宽；

30 网络侧设备覆盖下的终端分布特性。

在一些实施例中，所述场景类型信息包括以下至少之一：

视距(line of sight, LOS)、非视距(non line of sight, NLOS)、室外(outdoor)、室内(indoor)、城区微小区(urban microcell, UMi)、城区宏小区(urban macrocell, UMa)、郊区/农村宏小区(rural macrocell, RMa)、郊区/农村微小区(rural microcell, RMi)、
35 高速、低速。

在一些实施例中，第二信息可以是第一设备确定的，例如，第一设备根据第一信息和第一设备上存储的知识信息确定的，或者，根据第一信息和调用的知识信息确定的，例如第一设备可以使用调用接口信息从目标设备获取知识信息，进一步根据该知识信息，或者第一信息和该知识信息确定第二信息。

40 在另一些实施例中，第二信息也可以是使用调用接口信息所需访问的目标设备确定。可选地，目标设备确定第二信息后，可以直接发送给第三设备，或者，也可以通过其他设备发送给第三设备，例如通过网络开发功能转发，或者，通过第一设备发送给第三设备。

以下，结合具体实施例，对第三设备获取第二信息的具体实现进行说明。

45 实施例 1：第一设备上存储有用于辅助第一模型进行推理的知识信息。

此情况下，在第一设备从第二设备接收第一信息之后，第一设备可以根据第一信息

结合该第一设备上存储的知识信息，确定第二信息。进一步地，第一设备可以将第二信息发送给第三设备，从而第三设备可以在使用第一模型进行推理时可以使用第二信息作为辅助信息，能够提升模型的推理性能。

实施例 2：第一设备上未存储用于辅助第一模型进行推理的知识信息，并且第一信息包括知识信息的调用接口信息。此情况下，如图 4 所示，可以通过如下步骤实现：

S211，第二设备向第一设备发送第一信息，第一信息包括知识信息的调用接口信息，该调用接口信息可以是第二设备期望使用的知识信息的调用接口信息。此时，知识信息的调用接口信息包括了调用接口名称和调用接口的输入数据。

S212，第一设备根据该知识信息的调用接口信息，向目标设备发送第一信息，该目标设备是使用调用接口信息所需访问的设备；

S213，目标设备根据第一信息确定第二信息；

S214，目标设备向第一设备发送第二信息；

S215，第一设备向第三设备发送第二信息。

应理解，第一设备和目标设备之间的第一信息和第二信息的交互可以通过其他节点（例如网络开放功能）转发，或者，直接进行信息的交互。

实施例 3：第一设备上未存储用于辅助第一模型进行推理的知识信息，并且第一信息不包括知识信息的调用接口信息。

此情况下，作为一个实施例，如图 5 所示，可以通过如下步骤实现：

S221，第二设备向第一设备发送第一信息，第一信息不包括知识信息的调用接口信息。

S222，第一设备根据第一信息，确定知识信息的调用接口信息。

例如，第一设备可以根据第一信息中的用于第一模型推理的需求信息，在多个调用接口中确定目标调用接口，从而得到该目标调用接口的调用接口名称、调用接口的输入数据格式、调用接口的输出数据格式、调用接口的输入数据等信息。

S223，第一设备向目标设备发送第三信息，第三信息包括所述知识信息的调用接口信息。其中，该目标设备是使用所述调用接口信息所需访问的设备。所述知识信息的调用接口信息包括了调用接口名称和调用接口的输入数据。例如，该调用接口名称根据第一信息中的需求信息或辅助信息或调用接口信息确定，调用接口的输入数据是基于所述第一信息中的第一模型的输入信息确定。

S224，目标设备根据第三信息，确定第二信息；

S225，目标设备向第一设备发送第二信息；

S226，第一设备向第三设备发送第二信息。

如，在第二设备未指示调用接口信息的情况下，第一设备可以自主确定知识信息的调用接口信息，进一步将该调用接口信息，或者调用接口信息和第一信息发送给目标设备，从而目标设备可以基于该调用接口信息获取相应的知识信息，从而得到第二信息。然后目标设备可以将确定的第二信息发送给第一设备，由第一设备将该第二信息发送给第三设备。

应理解，第一设备和目标设备之间的第三信息和第二信息的交互可以通过其他节点（例如网络开放功能）转发，或者，直接进行信息的交互。

此情况下，作为另一个实施例，如图 6 所示，可以通过如下步骤实现：

S231，第二设备向第一设备发送第一信息，第一信息不包括知识信息的调用接口信息。

S232，第一设备根据第一信息，确定知识信息的调用接口信息。

例如，第一设备可以根据第一信息中的用于第一模型推理的需求信息，在多个调用接口中确定目标调用接口，从而得到该目标调用接口的调用接口名称、调用接口的输入数据格式、调用接口的输出数据格式、调用接口的输入数据等信息。

S233, 第一设备向第三设备发送第四信息, 第四信息包括所述知识信息的调用接口信息, 或者, 第四信息包括第一信息和所述知识信息的调用接口信息。其中, 所述知识信息的调用接口信息包括目标调用接口的调用接口名称, 或者目标调用接口的调用接口名称和调用接口的输入数据格式, 或者目标调用接口的调用接口名称和调用接口的输入数据格式和调用接口的输出数据格式。

S234, 第三设备根据第四信息向目标设备发送第七信息, 其中, 第七信息包括所述知识信息的调用接口信息; 其中, 所述知识信息的调用接口信息包括了调用接口名称和调用接口的输入数据。例如, 该调用接口名称包含在第四信息的所述知识信息的调用接口信息中, 调用接口的输入数据是基于第四信息中的所述知识信息的调用接口的输入数据格式和第三设备可提供的所述第一模型的输入信息确定。或者, 该调用接口名称根据第一信息 (包含在第四信息中) 中的需求信息或辅助信息或调用接口信息确定, 调用接口的输入数据是基于所述第一信息 (包含在第四信息中) 中的第一模型的输入信息确定。

S235, 目标设备根据第七信息确定第二信息, 并向第三设备发送第二信息。

如, 在第二设备未指示调用接口信息的情况下, 第一设备可以自主确定知识信息的调用接口信息, 进一步将该调用接口信息, 或者该调用接口信息和第一信息发送给第三设备, 由第三设备自主向目标设备请求第二信息, 然后目标设备可以将确定的第二信息发送给第三设备。

应理解, 第三设备和目标设备之间的第七信息和第二信息的交互可以通过其他节点 (例如网络开放功能) 转发, 或者, 直接进行信息的交互。

在本申请一些实施例中, 所述方法 200 还包括以下步骤至少之一:

第一设备接收所述目标设备发送的第五信息;

第一设备根据所述第五信息, 登记至少一个调用接口信息;

第一设备向所述目标设备发送第六信息。

在一些实施例中, 所述第五信息包括但不限于以下至少之一:

调用接口名称, 调用接口参数列表, 调用接口描述, 调用接口的使用样例, 调用接口指令集合; 其中, 调用接口参数列表包括调用接口的输入参数和输出参数, 每个参数包括参数名称, 参数描述, 数据类型, 缺省值中的至少之一。调用接口描述用于说明调用接口的功能。

在一些实施例中, 所述第六信息包括但不限于以下至少之一:

是否成功登记调用接口, 拒绝登记调用接口, 已登记的调用接口列表, 失败登记的调用接口列表。

进一步地, 在后续第一信息不包括调用接口信息时, 第一设备可以根据第一信息在登记的调用接口信息中确定目标调用接口信息, 使用该目标调用接口信息获取相应的知识信息。

在本申请一些实施例中, 若所述第三设备为终端, 所述方法 200 还包括:

所述第三设备向网络侧设备发送第八信息和第一能力信息中的至少之一;

其中, 所述第三设备上部署所述第一模型, 所述第八信息用于指示所述第三设备基于所述第一模型进行推理所需的辅助信息, 所述第一能力信息用于指示所述第三设备基于所述第一模型进行推理的能力信息。

在一些实施例中, 该第八信息或第一能力信息可以用于网络侧设备确定向终端指示的第二信息。

在一些实施例中, 所述第八信息包括以下至少之一:

第三设备基于所述第一模型进行推理所需的知识信息所在的网元指示;

第三设备基于所述第一模型进行推理所需的知识信息指示;

第三设备基于所述第一模型进行推理所需的知识信息的调用接口指示;

所述第二信息的预处理方式指示;

所述第二信息的格式指示。

可选地，第三设备基于所述第一模型进行推理所需的知识信息所在的网元指示可以是网元标识。例如，若有多个网元用于存储知识信息，第三设备进行模型推理所需的知识存储在第一个网元上，则第三设备可以向网络侧设备指示第一个网元的网元标识。

5 在一些实施例中，第三设备基于所述第一模型进行推理所需的知识信息指示可以包括第三设备基于所述第一模型进行推理所需的知识图谱的指示信息和/或知识向量库的指示信息。例如，知识图谱的指示信息是一个或多个知识图谱标识，或者，知识图谱组标识、知识图谱文件地址等，或者显式指示一个或多个知识图谱。例如，知识向量库的指示信息用于指示一个或多个知识向量库标识、或者显式指示一个或多个知识向量库。

10 示例性地，所述知识图谱的指示信息包括以下至少之一：

知识图谱标识；

知识图谱文件地址。

15 在一些实施例中，第三设备基于所述第一模型进行推理所需的知识信息的调用接口指示可以显式指示一个或多个调用接口信息，或者，也可以隐式指示一个或多个调用接口标识，该调用接口标识对应一组调用接口信息。

在一些实施例中，所述第二信息的格式指示可以用于指示第二信息的模型输入格式。例如，该模型输入格式可以包括模型输入包括的内容（例如包括提示信息，包括提示信息和调用接口信息，包括提示信息和网络侧设备的配置信息等），以及第二信息的维度，量化精度，大小限制等。

20 在一些实施例中，所述第二信息的预处理方式指示可以用于指示对第二信息进行预处理的方式，例如对数值做偏置，对数值做缩放，对数值做多项式函数变化，对单个数值做多维映射等。

在一些实施例中，所述第一能力信息包括以下至少之一：

第三设备基于所述第一模型进行推理所支持的知识信息指示；

25 第三设备基于所述第一模型进行推理所支持的知识信息的调用接口指示；

支持的所述第二信息的预处理方式；

支持的所述第二信息的格式指示。

30 在一些实施例中，第三设备基于所述第一模型进行推理所支持的知识信息指示可以包括第三设备基于所述第一模型进行推理所支持的知识图谱的指示信息和/或知识向量库的指示信息。例如，知识图谱的指示信息是一个或多个知识图谱标识，或者，知识图谱组标识、知识图谱文件地址等，或者显式指示一个或多个知识图谱。例如，知识向量库的指示信息用于指示一个或多个知识向量库标识、或者显式指示一个或多个知识向量库。

35 在一些实施例中，第三设备基于所述第一模型进行推理所支持的知识信息的调用接口指示可以显式指示一个或多个调用接口信息，或者，也可以隐式指示一个或多个调用接口标识，该调用接口标识对应一组调用接口信息。

40 在一些实施例中，所述第三设备支持的所述第二信息的格式指示可以用于指示第三设备支持的所述第二信息的模型输入格式，例如该模型输入格式可以包括模式输入包括哪些内容（例如包括提示信息，包括提示信息和调用接口信息，包括提示信息和网络侧设备的配置信息等），以及这些信息的维度，量化精度，大小限制等。

在一些实施例中，所述第三设备支持的所述第二信息的预处理方式可以用于指示第三设备支持的处理方法。这些处理方法是网络侧设备对第二信息进行预处理的方式，例如对数值做偏置，对数值做缩放，对数值做多项式函数变化，对单个数值做多维映射等。

45 在一些实施例中，第三设备在模型注册（或者说，模型识别（model identification））过程中或能力上报过程中，向所述网络侧设备发送所述第八信息和/或所述第一能力信息。

应理解，第八信息和第一能力信息可以是通过一个信令上报给网络侧设备的，或者，也可以是通过不同的信令上报给网络侧设备的，本申请对此不作限定。

以下，结合图 7 至图 14 所示具体实施例，说明本申请提供的通信系统中的模型推理方法。

5 实施例 1：推理设备是终端

如图 7 所示，第一设备为接入网设备或核心网功能，第二设备和第三设备同一设备，并且为终端，该终端上部署第一模型。该推理方法可以包括如下步骤：

步骤 1：终端向接入网设备或核心网设备发送第一信息，用于请求获取第二信息。

步骤 2：接入网设备或核心网设备根据第一信息向终端发送第二信息。

10 从而推理设备可以获知该第二信息，进一步地，推理设备可以使用该第二信息辅助进行模型的推理，能够提升模型的推理性能。

实施例 2：推理设备是接入网设备，例如基站。

例如，在图 8 的示例中，第一设备为核心网功能，第二设备和第三设备为同一设备，并且为接入网设备，该接入网设备上部署第一模型。

15 如图 8 所示，该推理方法可以包括如下步骤：

步骤 1：接入网设备向核心网设备发送第一信息，用于请求获取第二信息。

步骤 2：核心网设备根据第一信息向接入网设备发送第二信息。

从而推理设备可以获知该第二信息，进一步地，推理设备可以使用该第二信息辅助进行模型的推理，能够提升模型的推理性能。

20 又例如，在图 9 的示例中，第一设备为核心网功能，第二设备为终端，第三设备为接入网设备，该接入网设备上部署第一模型，如推理设备为接入网设备，例如基站。

如图 9 所示，该推理方法可以包括如下步骤：

步骤 1：终端向核心网功能发送第一信息，用于请求获取第二信息。

可选地，在步骤 1 中，终端还可以将第一信息发送给接入网设备。

25 可选地，终端发给接入网设备的第一信息与发送给核心网功能的第一信息可以不同。

例如，终端发送给核心网的第一信息可以包括用于所述第一模型进行推理的辅助信息和所述第一模型的输入信息；而终端发送给接入网设备的第一信息仅包括所述第一模型的输入信息。此外，为了让核心网设备知道将第二信息发给谁，第一信息还可以包括接收设备指示，用于指示第二信息的接收设备，如第三设备。

30 步骤 2：核心网设备根据第一信息向接入网设备发送第二信息。

此时，接入网设备用于第一模型的输入包括两部分，一部分来自终端发送的第一信息，一部分来自核心网发送的第二信息。因此，接入网设备需要将两部分信息进行同步从而生成完整的第一模型的输入信息。从而防止将样本 1 的第一信息和样本 2 的第二信息进行合并而生成错误的样本。因此，此时第一信息和第二信息中需要包含样本指示，
35 例如样本标识。

可选地，核心网设备在向接入网设备发送第二信息时，也可以将该第一信息一起发送给接入网设备。从而推理设备可以获知该第一信息和第二信息，进一步地，推理设备可以使用该第一信息和第二信息辅助进行模型的推理，能够提升模型的推理性能。

40 又例如，在图 10 的示例中，第一设备为核心网功能，第二设备为终端，第三设备为接入网设备，该接入网设备上部署第一模型，如推理设备为接入网设备，例如基站。

如图 10 所示，该推理方法可以包括如下步骤：

步骤 1：终端向核心网设备发送第一信息，用于请求获取第二信息。

步骤 2：核心网设备根据第一信息向终端发送第二信息。

步骤 3：终端向接入网设备发送第二信息。

45 可选地，终端还可以将第一信息一起发送给接入网设备。

如，在该示例中，第二信息不是由知识存储设备直接发送给推理设备的，而是由需

求发起设备转发给推理设备的。

从而推理设备可以获知该第一信息和第二信息，进一步地，推理设备可以使用该第一信息和第二信息辅助进行模型的推理，能够提升模型的推理性能。

又例如，在图 11 的示例中，第一设备为核心网功能，第二设备为终端，第三设备为接入网设备，该接入网设备上部署第一模型，如推理设备为接入网设备，例如基站。

如图 11 所示，该推理方法可以包括如下步骤：

步骤 1：终端向核心网设备发送第一信息，用于请求获取第二信息。

其中，第一信息包括如下至少之一：

接收设备指示，用于指示第二信息的接收设备。

用于所述第一模型进行推理的辅助信息；

用于所述第一模型进行推理的需求信息；

所述第一模型的输入信息。步骤 2：核心网设备根据第一信息向接入网设备发送第一信息和第二信息。

从而推理设备可以获知该第一信息和第二信息，进一步地，推理设备可以使用该第一信息和第二信息辅助进行模型的推理，能够提升模型的推理性能。

实施例 3：推理设备是核心网功能。

例如，在图 12 的示例中，第一设备为核心网功能，第二设备为终端或接入网设备或核心网设备，第三设备为核心网功能。

如图 12 所示，该推理方法可以包括如下步骤：

步骤 1：第二设备向第一设备发送第一信息，用于请求获取第二信息。

其中，第一信息包括如下至少之一：

接收设备指示，用于指示第二信息的接收设备；

用于所述第一模型进行推理的辅助信息；

用于所述第一模型进行推理的需求信息；

所述第一模型的输入信息。

可选地，在步骤 1 中，第二设备还可以将第一信息也发送给第三设备，用于辅助第三设备进行模型推理。此时，第三设备用于第一模型的输入包括两部分，一部分来自第二设备发送的第一信息，一部分来自第一设备发送的第二信息。因此，第三设备需要将两部分信息进行同步从而生成完整的第一模型的输入信息。从而防止将样本 1 的第一信息和样本 2 的第二信息进行合并而生成错误的样本。因此，此时第一信息和第二信息中需要包含样本指示，例如样本标识。

步骤 2：第一设备根据第一信息向第三设备发送第二信息。

从而推理设备可以获知该第一信息和第二信息，进一步地，推理设备可以使用该第一信息和第二信息辅助进行模型的推理，能够提升模型的推理性能。

又例如，在图 13 的示例中，第一设备为核心网功能，第二设备为终端或接入网设备或核心网设备，第三设备为核心网功能。

如图 13 所示，该推理方法可以包括如下步骤：

步骤 1：第二设备向第一设备发送第一信息，用于请求获取第二信息。

步骤 2：第一设备根据第一信息向第二设备发送第二信息。

步骤 3：第二设备向第三设备发送第二信息。可选地，还可以将第一信息也发送给第三设备。

从而推理设备可以获知该第二信息，或者第一信息和第二信息，进一步地，推理设备可以使用该第二信息，或者第一信息和第二信息辅助进行模型的推理，能够提升模型的推理性能。

又例如，在图 14 的示例中，第一设备为核心网功能，第二设备为终端或接入网设备或核心网设备，第三设备为核心网功能。

如图 14 所示, 该推理方法可以包括如下步骤:

步骤 1: 第二设备向第一设备发送第一信息, 用于请求获取第二信息。

其中, 第一信息包括如下至少之一:

接收设备指示, 用于指示第二信息的接收设备;

5 用于所述第一模型进行推理的辅助信息;

用于所述第一模型进行推理的需求信息;

所述第一模型的输入信息。

步骤 2: 第一设备根据第一信息向第三设备发送第一信息和第二信息。

10 从而推理设备可以获知该第二信息, 或者第一信息和第二信息, 进一步地, 推理设备可以使用该第二信息, 或者第一信息和第二信息辅助进行模型的推理, 能够提升模型的推理性能。

应理解, 在本申请实施例中, 在推理设备获知第一信息和第二信息的情况下, 该第一信息和第二信息中包括样本指示, 用于避免推理设备使用不同样本的第一信息和第二信息进行推理, 影响推理结果。

15 综上, 在本申请实施例中, 通过向存储有知识信息的设备获取第二信息, 进一步使用模型进行推理时, 可以获得更多的信息用作模型的输入, 从而能够提升模型的推理性能。

20 上文结合图 2 至图 14, 详细描述了本申请的方法实施例, 下文结合图 15 至图 21, 详细描述本申请的装置实施例, 应理解, 装置实施例与方法实施例相互对应, 类似的描述可以参照方法实施例。

本申请实施例提供的通信系统中的模型推理方法, 执行主体可以为通信装置。本申请实施例中以通信装置执行通信系统中的模型推理方法为例, 说明本申请实施例提供的通信装置。

25 图 15 示出了根据本申请实施例的通信装置 500 的示意性框图。如图 15 所示, 该装置 500 包括:

通信单元 510, 用于接收第二设备发送的第一信息, 其中, 所述通信装置 500 上存储有知识信息和知识信息的调用接口信息中的至少之一; 其中, 所述知识信息包括以下至少之一: 知识图谱、知识向量库; 以及

30 向第三设备发送第二信息, 所述第三设备上部署第一模型, 所述第二信息是根据所述第一信息和所述通信装置 500 上存储的或调用的知识信息确定的;

其中, 所述第一信息包括以下至少之一:

用于所述第一模型进行推理的辅助信息;

用于所述第一模型进行推理的需求信息;

所述第一模型的输入信息;

35 知识信息的调用接口信息;

所述第二信息的接收设备指示;

所述第一模型的输入信息对应的样本指示。

在一些实施例中, 所述辅助信息包括以下至少之一:

40 人工智能 AI 模型的标识信息、AI 模型的功能指示、AI 模型的特性指示、目标知识图谱指示、小区标识、带宽部分 BWP 标识。

在一些实施例中, 所述需求信息包括以下至少之一:

AI 模型所需适配的目标任务指示;

AI 模型的业务体验质量 QoE 要求;

AI 模型的业务处理时延;

45 AI 模型的业务处理准确度;

AI 模型的业务计算量;

AI 模型的数据处理规模。

在一些实施例中，所述第二信息包括以下至少之一：

隐藏变量，所述隐藏变量是对所述第一模型的输入信息进行映射后的特征向量；
提示信息；

5 网络侧设备的配置信息；

场景类型信息；

所述第一模型的输入信息对应的样本指示。

在一些实施例中，所述提示信息包括以下至少之一：

与所述第一模型适配的目标任务相关的思维链；

10 与所述目标任务相关的提示文字；

与所述目标任务相关的提示文件；

与所述目标任务相关的提示代码；

与所述目标任务相关的思维链标识；

与所述目标任务相关的提示文字标识；

15 与所述目标任务相关的提示文件标识；

与所述目标任务相关的提示代码标识；

与所述目标任务相关的提示文件地址。

在一些实施例中，所述网络侧设备的配置信息包括以下至少之一：

网络侧设备的发送天线个数；

20 网络侧设备的发送波束个数；

网络侧设备的天线端口个数；

网络侧设备的发送功率；

网络侧设备的天线增益；

网络侧设备的波束 3dB 带宽；

25 网络侧设备的间距；

网络侧设备的频率；

系统带宽；

网络侧设备覆盖下的终端分布特性。

在一些实施例中，所述场景类型信息包括以下至少之一：

30 视距 LOS、非视距 NLOS、室内、室外、城区微小区、城区宏小区、郊区微小区、
高速、低速。

在一些实施例中，所述通信装置 500 还包括：

处理单元，用于根据所述第一信息和所述通信装置 500 上存储的知识信息确定所述
第二信息。

35 在一些实施例中，所述通信单元 510 还用于：

从目标设备接收所述第二信息，所述目标设备是使用所述调用接口信息所访问的设
备。

在一些实施例中，所述第一信息包括知识信息的调用接口信息，所述通信装置 500
还包括：处理单元，用于根据所述第一信息和所述通信装置 500 上存储的知识信息确定
40 所述第二信息。

在一些实施例中，所述通信单元 510 还用于：根据所述知识信息的调用接口信息，
向目标设备发送所述第一信息，其中，所述目标设备是使用所述调用接口信息所访问的
设备；以及从所述目标设备接收所述第二信息。

45 在一些实施例中，所述第一信息不包括知识信息的调用接口信息，所述通信装置 500
还包括：处理单元，用于根据所述第一信息，确定所述知识信息的调用接口信息；

所述通信装置 500 还用于：向目标设备发送第三信息，其中，所述目标设备是使用

所述调用接口信息所访问的设备，所述第三信息包括所述知识信息的调用接口信息；以及，从所述目标设备接收所述第二信息。

在一些实施例中，所述第一信息不包括知识信息的调用接口信息，所述通信装置 500 还包括：处理单元，用于根据所述第一信息，确定所述知识信息的调用接口信息；

5 所述通信单元 510 还用于：向所述第三设备发送第四信息，所述第四信息包括所述知识信息的调用接口信息，或者，所述第四信息包括所述第一信息和所述知识信息的调用接口信息，所述第四信息用于所述第三设备向目标设备请求获取所述第二信息，所述目标设备是使用所述调用接口信息所访问的设备。

在一些实施例中，所述通信单元 510 还用于：

10 从所述目标设备接收第五信息；

所述通信装置 500 还包括处理单元，用于根据所述第五信息，登记至少一个调用接口信息；

所述通信单元 510 还用于：向所述目标设备发送第六信息；

15 其中，所述第五信息包括以下至少之一：调用接口名称，调用接口参数列表，调用接口描述，调用接口的使用样例，调用接口指令集合；

所述第六信息包括以下至少之一：

是否成功登记调用接口，拒绝登记调用接口，已登记的调用接口列表，失败登记的调用接口列表。

在一些实施例中，所述通信单元 510 还用于：

20 向所述第三设备发送所述第一信息。

在一些实施例中，所述通信装置 500 为接入网设备或核心网功能，所述第二设备和所述第三设备为终端；或者

所述通信装置 500 为核心网功能，所述第二设备为终端或接入网设备，所述第三设备为接入网设备；或者

25 所述通信装置 500 为核心网功能，所述第二设备为终端或接入网设备或核心网功能，所述第三设备为核心网功能。

在一些实施例中，所述通信装置 500 为接入网设备，所述第二设备为终端，所述第一信息携带在以下至少一种信令中：层 1 信令、层 2 信令、层 3 信令、AI 层信令、数据面信令；或者

30 所述通信装置 500 为核心网功能，所述第三设备为终端，所述第一信息携带在以下至少一种信令中：非接入层 NAS 信令、AI 层信令、数据面信令。

在一些实施例中，所述调用接口信息包括如下至少之一：

调用接口名称、调用接口的输入数据格式、调用接口的输出数据格式、调用接口的输入数据。

35 可选地，在一些实施例中，上述通信单元可以是通信接口或收发器，或者是通信芯片或者片上系统的输入输出接口。

应理解，根据本申请实施例的通信装置 500 可对应于本申请方法实施例中的第一设备，并且通信装置 500 中的各个单元的上述和其它操作和/或功能分别为了实现图 2 至图 14 中所示方法实施例中第一设备的相应流程，并达到相同的技术效果，为避免重复，这里不再赘述。

40 图 16 示出了根据本申请实施例的通信装置 600 的示意性框图。如图 16 所示，该装置 600 包括：

通信单元 610，用于向第一设备发送第一信息，其中，所述第一设备上存储有知识信息和知识信息的调用接口信息中的至少之一，其中，所述知识信息包括以下至少之一：

45 知识图谱、知识向量库；

其中，所述第一信息包括以下至少之一：

用于第一模型进行推理的辅助信息；
用于第一模型进行推理的需求信息；
第一模型的输入信息；
知识信息的调用接口信息；

5 所述第二信息的接收设备指示；
所述第一模型的输入信息对应的样本指示。

在一些实施例中，所述辅助信息包括以下至少之一：

人工智能 AI 模型的标识信息、AI 模型的功能指示、AI 模型的特性指示、目标知识
图谱指示、小区标识、带宽部分 BWP 标识。

10 在一些实施例中，所述需求信息包括以下至少之一：

AI 模型所需适配的目标任务指示；
AI 模型的业务体验质量 QoE 要求；
AI 模型的业务处理时延；
AI 模型的业务处理准确度；

15 AI 模型的业务计算量；
AI 模型的数据处理规模。

在一些实施例中，所述第二信息包括以下至少之一：

隐藏变量，所述隐藏变量是对所述第一模型的输入信息进行映射后的特征向量；
提示信息；

20 网络侧设备的配置信息；
场景类型信息；
所述第一模型的输入信息对应的样本指示。

在一些实施例中，所述提示信息包括以下至少之一：

与所述第一模型适配的目标任务相关的思维链；

25 与所述目标任务相关的提示文字；
与所述目标任务相关的提示文件；
与所述目标任务相关的提示代码；
与所述目标任务相关的思维链标识；
与所述目标任务相关的提示文字标识；
30 与所述目标任务相关的提示文件标识；
与所述目标任务相关的提示代码标识；
与所述目标任务相关的提示文件地址。

在一些实施例中，所述网络侧设备的配置信息包括以下至少之一：

网络侧设备的发送天线个数；
35 网络侧设备的发送波束个数；
网络侧设备的天线端口个数；
网络侧设备的发送功率；
网络侧设备的天线增益；
网络侧设备的波束 3dB 带宽；

40 网络侧设备的间距；
网络侧设备的频率；
系统带宽；

网络侧设备覆盖下的终端分布特性。

在一些实施例中，所述场景类型信息包括以下至少之一：

45 视距 LOS、非视距 NLOS、室内、室外、城区微小区、城区宏小区、郊区微小区、
高速、低速。

在一些实施例中，所述调用接口信息包括如下至少之一：

调用接口名称、调用接口的输入数据格式、调用接口的输出数据格式、调用接口的输入数据。

在一些实施例中，所述通信装置 600 为终端，所述通信单元 610 具体用于：

- 5 在发生以下事件中的至少之一时，向所述第一设备发送所述第一信息：
接收到第一指示，所述第一指示用于指示激活或切换至所述第一模型；
所述通信装置 600 确定激活或切换至所述第一模型；
所述通信装置 600 确定所述第一模型不满足需求；
接收到第二指示，所述第二指示用于指示所述第一模型不满足需求。

- 10 在一些实施例中，所述通信单元 610 还用于：
向第三设备发送所述第一信息，所述第三设备上部署所述第一模型。

可选地，在一些实施例中，上述通信单元可以是通信接口或收发器，或者是通信芯片或者片上系统的输入输出接口。

- 15 应理解，根据本申请实施例的通信装置 600 可对应于本申请方法实施例中的第二设备，并且通信装置 600 中的各个单元的上述和其它操作和/或功能分别为了实现图 2 至图 14 中所示方法实施例中第二设备的相应流程，并达到相同的技术效果，为避免重复，这里不再赘述。

图 17 示出了根据本申请实施例的通信装置 700 的示意性框图。如图 17 所示，该装置 700 包括：

- 20 通信单元 710，用于从第一设备或目标设备获取第二信息，其中，所述通信装置 700 上部署第一模型，所述第一设备或所述目标设备上存储有知识信息和知识信息的调用接口信息中的至少之一，其中，所述目标设备是使用所述调用接口信息所访问的设备，所述知识信息包括以下至少之一：知识图谱、知识向量库；

其中，所述第二信息根据第一信息和所述第一设备上存储或调用的知识信息确定；

- 25 其中，所述第一信息包括以下至少之一：

用于所述第一模型进行推理的辅助信息；

用于所述第一模型进行推理的需求信息；

所述第一模型的输入信息；

知识信息的调用接口信息；

- 30 所述第二信息的接收设备指示；

所述第一模型的输入信息对应的样本指示。

在一些实施例中，所述辅助信息包括以下至少之一：

人工智能 AI 模型的标识信息、AI 模型的功能指示、AI 模型的特性指示、目标知识图谱指示、小区标识、带宽部分 BWP 标识。

- 35 在一些实施例中，所述需求信息包括以下至少之一：

AI 模型所需适配的目标任务指示；

AI 模型的业务体验质量 QoE 要求；

AI 模型的业务处理时延；

AI 模型的业务处理准确度；

- 40 AI 模型的业务计算量；

AI 模型的数据处理规模。

在一些实施例中，所述第二信息包括以下至少之一：

隐藏变量，所述隐藏变量是对所述第一模型的输入信息进行映射后的特征向量；

提示信息；

- 45 网络侧设备的配置信息；

场景类型信息；

所述第一模型的输入信息对应的样本指示。

在一些实施例中，所述提示信息包括以下至少之一：

与所述第一模型适配的目标任务相关的思维链；

与所述目标任务相关的提示文字；

5 与所述目标任务相关的提示文件；

与所述目标任务相关的提示代码；

与所述目标任务相关的思维链标识；

与所述目标任务相关的提示文字标识；

与所述目标任务相关的提示文件标识；

10 与所述目标任务相关的提示代码标识；

与所述目标任务相关的提示文件地址。

在一些实施例中，所述网络侧设备的配置信息包括以下至少之一：

网络侧设备的发送天线个数；

网络侧设备的发送波束个数；

15 网络侧设备的天线端口个数；

网络侧设备的发送功率；

网络侧设备的天线增益；

网络侧设备的波束 3dB 带宽；

网络侧设备的间距；

20 网络侧设备的频率；

系统带宽；

网络侧设备覆盖下的终端分布特性。

在一些实施例中，所述场景类型信息包括以下至少之一：

25 视距 LOS、非视距 NLOS、室内、室外、城区微小区、城区宏小区、郊区微小区、
高速、低速。

在一些实施例中，所述通信单元 710 还用于：

从所述第一设备接收第四信息，所述第四信息包括所述知识信息的调用接口信息，
或者，所述第四信息包括所述第一信息和所述知识信息的调用接口信息；

向所述目标设备发送第七信息，所述第七信息包括所述知识信息的调用接口信息；

30 从所述目标设备接收所述第二信息。

在一些实施例中，所述通信装置 700 为终端，所述通信单元 710 还用于：

向网络侧设备发送第八信息和第一能力信息中的至少之一，其中，所述通信装置 700
上部署所述第一模型，所述第八信息用于指示所述通信装置 700 基于所述第一模型进行
推理所需的辅助信息，所述第一能力信息用于指示所述通信装置 700 基于所述第一模型
35 进行推理的能力信息。

在一些实施例中，所述第八信息包括以下至少之一：

通信装置 700 基于所述第一模型进行推理所需的知识信息所在的网元指示；

通信装置 700 基于所述第一模型进行推理所需的知识信息指示；

通信装置 700 基于所述第一模型进行推理所需的知识信息的调用接口指示；

40 所述第二信息的格式指示。

在一些实施例中，所述第一能力信息包括以下至少之一：

通信装置 700 基于所述第一模型进行推理所支持的知识信息指示；

通信装置 700 基于所述第一模型进行推理所支持的知识信息的调用接口指示；

支持的所述第二信息的格式指示。

45 在一些实施例中，所述通信单元 710 还用于：

在模型注册过程中或能力上报过程中，向所述网络侧设备发送所述第八信息和所述

第一能力信息中的至少之一。

在一些实施例中，所述通信单元 710 还用于：

从第二设备接收所述第一信息，所述通信装置 700 上部署所述第一模型。

在一些实施例中，所述通信装置 700 还包括：

5 处理单元，用于利用所述第一模型基于所述第二信息和所述第一模型的输入信息进行推理。

在一些实施例中，所述调用接口信息包括如下至少之一：

调用接口名称、调用接口的输入数据格式、调用接口的输出数据格式、调用接口的输入数据。

10 可选地，在一些实施例中，上述通信单元可以是通信接口或收发器，或者是通信芯片或者片上系统的输入输出接口。

应理解，根据本申请实施例的通信装置 700 可对应于本申请方法实施例中的第三设备，并且通信装置 700 中的各个单元的上述和其它操作和/或功能分别为了实现图 2 至图 14 中所示方法实施例中第三设备的相应流程，并达到相同的技术效果，为避免重复，这里不再赘述。

15 在一些实施例中，本申请实施例中的装置 500、装置 600、装置 700 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。

20 如图 18 所示，本申请实施例还提供一种通信设备 1000，包括处理器 1001 和存储器 1002，存储器 1002 上存储有可在所述处理器 1001 上运行的程序或指令，例如，该通信设备 1000 为第一设备时，该程序或指令被处理器 1001 执行时实现上述推理方法实施例中由第一设备执行的步骤，且能达到相同的技术效果。例如，该通信设备 1000 为第二设备时，该程序或指令被处理器 1001 执行时实现上述推理方法实施例中由第二设备执行的步骤，且能达到相同的技术效果。例如，该通信设备 1000 为第三设备时，该程序或指令被处理器 1001 执行时实现上述推理方法实施例中由第三设备执行的步骤，且能达到相同的技术效果。为避免重复，这里不再赘述。

30 本申请实施例还提供一种终端，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，实现如图 2 至图 14 所示方法实施例中的步骤。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。具体地，图 19 为实现本申请实施例的一种终端的硬件结构示意图。

35 该终端 1100 包括但不限于：射频单元 1101、网络模块 1102、音频输出单元 1103、输入单元 1104、传感器 1105、显示单元 1106、用户输入单元 1107、接口单元 1108、存储器 1109 以及处理器 1110 等中的至少部分部件。

40 本领域技术人员可以理解，终端 1100 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 1110 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 19 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。

45 应理解的是，本申请实施例中，输入单元 1104 可以包括图形处理单元（Graphics Processing Unit, GPU）11041 和麦克风 11042，图形处理器 11041 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 1106 可包括显示面板 11061，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 11061。用户输入单元 1107 包括触控面板 11071 以及其他输入设备

11072 中的至少一种。触控面板 11071, 也称为触摸屏。触控面板 11071 可包括触摸检测装置和触摸控制器两个部分。其他输入设备 11072 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 1101 接收来自网络侧设备的下行数据后, 可以传输给处理器 1110 进行处理; 另外, 射频单元 1101 可以向网络侧设备发送上行数据。通常, 射频单元 1101 包括但不限于天线、放大器、收发信机、耦合器、低噪声放大器、双工器等。

存储器 1109 可用于存储软件程序或指令以及各种数据。存储器 1109 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 1109 可以包括易失性存储器或非易失性存储器。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 1109 包括但不限于这些和任意其它适合类型的存储器。

处理器 1110 可包括一个或多个处理单元; 可选的, 处理器 1110 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 1110 中。

可以理解, 本实施例中提及的各实现方式的实现过程可以参照图 2 至图 14 所示方法实施例的相关描述, 并达到相同或相应的技术效果, 为避免重复, 在此不再赘述。

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, 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, 实现如图 2 至图 14 所示所示的方法实施例的步骤。该网络侧设备实施例与上述接入网设备侧或核心网功能侧方法实施例对应, 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网络侧设备实施例中, 且能达到相同的技术效果。

具体地, 本申请实施例还提供了一种网络侧设备。如图 20 所示, 该网络侧设备 1200 包括: 天线 1201、射频装置 1202、基带装置 1203、处理器 1204 和存储器 1205。天线 1201 与射频装置 1202 连接。在上行方向上, 射频装置 1202 通过天线 1201 接收信息, 将接收的信息发送给基带装置 1203 进行处理。在下行方向上, 基带装置 1203 对要发送的信息进行处理, 并发送给射频装置 1202, 射频装置 1202 对收到的信息进行处理后经过天线 1201 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 1203 中实现, 该基带装置 1203 包括基带处理器。

基带装置 1203 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 20 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 1205 连接, 以调用存储器 1205 中的程序, 执行以上方法实施例所示的网络设备操作。

该网络侧设备还可以包括网络接口 1206, 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 1200 还包括: 存储在存储器 1205 上并可在处理器 1204 上运行的指令或程序, 处理器 1204 调用存储器 1205 中的指令或程序执行图 15

至图 17 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。

具体地，本申请实施例还提供了一种网络侧设备。如图 21 所示，该网络侧设备 1300 包括：处理器 1301、网络接口 1302 和存储器 1303。其中，网络接口 1302 例如为通用公共无线接口（common public radio interface, CPRI）。

5 具体地，本申请实施例的网络侧设备 1300 还包括：存储在存储器 1303 上并可在处理器 1301 上运行的指令或程序，处理器 1301 调用存储器 1303 中的指令或程序执行图 15 至图 17 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。

10 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通信系统中的模型推理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通信装置、通信设备、终端或网络侧设备中的处理器。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。在一些示例中，可读存储介质可以是非瞬态的可读存储介质。

15 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，实现上述通信系统中的模型推理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。

20 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通信系统中的模型推理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。

25 本申请实施例还提供了一种通信系统，包括：第一设备、第二设备和第三设备，所述第一设备可用于执行如上所述的通信系统中模型推理方法中由第一设备执行的步骤，第二设备可用于执行如上所述的通信系统中的模型推理方法中由第二设备执行的步骤，第三设备可用于执行如上所述的通信系统中的模型推理方法中由第三设备执行的步骤。

30 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，
35 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。

40 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计算机软件产品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。该计算机软件产品存储在存储介质（如 ROM、RAM、磁碟、光盘等）中，包括若干指令，用以使得终端或者网络侧设备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，还可做出很多形式的实施方式，这些实施方式均属于本申请的保护之内。

权利要求书

1、一种通信系统中的模型推理方法，包括：

第一设备接收第二设备发送的第一信息，其中，所述第一设备上存储有知识信息和知识信息的调用接口信息中的至少之一；其中，所述知识信息包括以下至少之一：知识图谱、知识向量库；

所述第一设备向第三设备发送第二信息，所述第三设备上部署第一模型，所述第二信息是根据所述第一信息和所述第一设备上存储的或调用的知识信息确定的；

其中，所述第一信息包括以下至少之一：

用于所述第一模型进行推理的辅助信息；

用于所述第一模型进行推理的需求信息；

所述第一模型的输入信息；

知识信息的调用接口信息；

所述第二信息的接收设备指示；

所述第一模型的输入信息对应的样本指示。

2、根据权利要求1所述的方法，其中，所述辅助信息包括以下至少之一：

人工智能 AI 模型的标识信息、AI 模型的功能指示、AI 模型的特性指示、目标知识图谱指示、小区标识、带宽部分 BWP 标识。

3、根据权利要求1或2所述的方法，其中，所述需求信息包括以下至少之一：

AI 模型所需适配的目标任务指示；

AI 模型的业务体验质量 QoE 要求；

AI 模型的业务处理时延；

AI 模型的业务处理准确度；

AI 模型的业务计算量；

AI 模型的数据处理规模。

4、根据权利要求1-3中任一项所述的方法，其中，所述第二信息包括以下至少之一：

隐藏变量，所述隐藏变量是对所述第一模型的输入信息进行映射后的特征向量；

提示信息；

网络侧设备的配置信息；

场景类型信息；

所述第一模型的输入信息对应的样本指示；

其中，所述提示信息包括以下至少之一：

与所述第一模型适配的目标任务相关的思维链；

与所述目标任务相关的提示文字；

与所述目标任务相关的提示文件；

与所述目标任务相关的提示代码；

与所述目标任务相关的思维链标识；

与所述目标任务相关的提示文字标识；

与所述目标任务相关的提示文件标识；

与所述目标任务相关的提示代码标识；

与所述目标任务相关的提示文件地址；

其中，所述网络侧设备的配置信息包括以下至少之一：

网络侧设备的发送天线个数；

网络侧设备的发送波束个数；

网络侧设备的天线端口个数；

网络侧设备的发送功率；

网络侧设备的天线增益；

网络侧设备的波束 3dB 带宽；

网络侧设备的间距；

网络侧设备的频率；

系统带宽；

5 网络侧设备覆盖下的终端分布特性；

其中，所述场景类型信息包括以下至少之一：

视距 LOS、非视距 NLOS、室内、室外、城区微小区、城区宏小区、郊区微小区、高速、低速。

5、根据权利要求 1-4 中任一项所述的方法，其中，所述方法还包括：

10 所述第一设备根据所述第一信息和所述第一设备上存储的知识信息确定所述第二信息；

或者，

所述第一设备从目标设备接收所述第二信息，所述目标设备是使用所述调用接口信息所访问的设备。

15 6、根据权利要求 5 所述的方法，其中，所述方法还包括：

所述第一设备接收所述目标设备发送的第五信息；

所述第一设备根据所述第五信息，登记至少一个调用接口信息；

所述第一设备向所述目标设备发送第六信息；

20 其中，所述第五信息包括以下至少之一：调用接口名称，调用接口参数列表，调用接口描述，调用接口的使用样例，调用接口指令集合；

所述第六信息包括以下至少之一：

是否成功登记调用接口，拒绝登记调用接口，已登记的调用接口列表，失败登记的调用接口列表。

7、根据权利要求 1-6 中任一项所述的方法，其中，所述方法还包括：

25 所述第一设备向所述第三设备发送所述第一信息。

8、根据权利要求 1-7 中任一项所述的方法，其中，所述调用接口信息包括如下至少之一：

调用接口名称、调用接口的输入数据格式、调用接口的输出数据格式、调用接口的输入数据。

30 9、一种通信系统中的模型推理方法，包括：

第二设备向第一设备发送第一信息，其中，所述第一设备上存储有知识信息和知识信息的调用接口信息中的至少之一，其中，所述知识信息包括以下至少之一：知识图谱、知识向量库；

其中，所述第一信息包括以下至少之一：

35 用于第一模型进行推理的辅助信息；

用于第一模型进行推理的需求信息；

第一模型的输入信息；

知识信息的调用接口信息；

第二信息的接收设备指示；

40 所述第一模型的输入信息对应的样本指示。

10、根据权利要求 9 所述的方法，其中，所述第二设备为终端，所述第二设备向第一设备发送第一信息，包括：

在发生以下事件中的至少之一时，向所述第一设备发送所述第一信息：

接收到第一指示，所述第一指示用于指示激活或切换至所述第一模型；

45 所述第二设备确定激活或切换至所述第一模型；

所述第二设备确定所述第一模型不满足需求；

接收到第二指示，所述第二指示用于指示所述第一模型不满足需求。

11、根据权利要求 9 或 10 所述的方法，其中，所述方法还包括：

所述第二设备向第三设备发送所述第一信息，所述第三设备上部署所述第一模型。

12、一种通信系统中的模型推理方法，包括：

5 第三设备从第一设备或目标设备获取第二信息，其中，所述第三设备上部署第一模型，所述第一设备或所述目标设备上存储有知识信息和知识信息的调用接口信息中的至少之一，其中，所述目标设备是使用所述调用接口信息所访问的设备，所述知识信息包括以下至少之一：知识图谱、知识向量库；

其中，所述第二信息根据第一信息和所述第一设备上存储或调用的知识信息确定；

10 其中，所述第一信息包括以下至少之一：

用于所述第一模型进行推理的辅助信息；

用于所述第一模型进行推理的需求信息；

所述第一模型的输入信息；

知识信息的调用接口信息；

15 所述第二信息的接收设备指示；

所述第一模型的输入信息对应的样本指示。

13、根据权利要求 12 所述的方法，其中，所述第三设备从所述目标设备获取第二信息，包括：

20 所述第三设备接收所述第一设备发送的第四信息，所述第四信息包括所述知识信息的调用接口信息，或者，所述第四信息包括所述第一信息和所述知识信息的调用接口信息；

所述第三设备向所述目标设备发送第七信息，所述第七信息包括所述知识信息的调用接口信息；

所述第三设备接收所述目标设备发送的所述第二信息。

25 14、根据权利要求 12 或 13 所述的方法，其中，所述第三设备为终端，所述方法还包括：

30 所述第三设备向网络侧设备发送第八信息和第一能力信息中的至少之一，其中，所述第三设备上部署所述第一模型，所述第八信息用于指示所述第三设备基于所述第一模型进行推理所需的辅助信息，所述第一能力信息用于指示所述第三设备基于所述第一模型进行推理的能力信息。

15、根据权利要求 14 所述的方法，其中，所述第八信息包括以下至少之一：

第三设备基于所述第一模型进行推理所需的知识信息所在的网元指示；

第三设备基于所述第一模型进行推理所需的知识信息指示；

第三设备基于所述第一模型进行推理所需的知识信息的调用接口指示；

35 所述第二信息的预处理方式指示；

所述第二信息的格式指示。

16、根据权利要求 14 或 15 所述的方法，其中，所述第一能力信息包括以下至少之一：

第三设备基于所述第一模型进行推理所支持的知识信息指示；

40 第三设备基于所述第一模型进行推理所支持的知识信息的调用接口指示；

支持的所述第二信息的预处理方式指示；

支持的所述第二信息的格式指示。

17、根据权利要求 14-16 中任一项所述的方法，其中，所述第三设备向网络侧设备发送第八信息和第一能力信息中的至少之一，包括：

45 所述第三设备在模型注册过程中或能力上报过程中，向所述网络侧设备发送所述第八信息和所述第一能力信息中的至少之一。

18、根据权利要求 12-17 中任一项所述的方法，其中，所述方法还包括：

所述第三设备接收第二设备发送的所述第一信息，所述第三设备上部署所述第一模型。

19、根据权利要求 12-18 中任一项所述的方法，其中，所述方法还包括：

5 所述第三设备利用所述第一模型基于所述第二信息和所述第一模型的输入信息进行推理。

20、根据权利要求 12-19 中任一项所述的方法，其中，所述调用接口信息包括如下至少之一：

10 调用接口名称、调用接口的输入数据格式、调用接口的输出数据格式、调用接口的输入数据。

21、一种通信装置，包括：

通信单元，用于接收第二设备发送的第一信息，其中，所述通信装置上存储有知识信息和知识信息的调用接口信息中的至少之一；其中，所述知识信息包括以下至少之一：知识图谱、知识向量库；以及

15 向第三设备发送第二信息，所述第三设备上部署第一模型，所述第二信息是根据所述第一信息和所述通信装置上存储的或调用的知识信息确定的；

其中，所述第一信息包括以下至少之一：

用于所述第一模型进行推理的辅助信息；

用于所述第一模型进行推理的需求信息；

20 所述第一模型的输入信息；

知识信息的调用接口信息；

所述第二信息的接收设备指示；

所述第一模型的输入信息对应的样本指示。

22、一种通信装置，包括：

25 通信单元，用于向第一设备发送第一信息，其中，所述第一设备上存储有知识信息和知识信息的调用接口信息中的至少之一，其中，所述知识信息包括以下至少之一：知识图谱、知识向量库；

其中，所述第一信息包括以下至少之一：

用于第一模型进行推理的辅助信息；

30 用于第一模型进行推理的需求信息；

第一模型的输入信息；

知识信息的调用接口信息；

第二信息的接收设备指示；

所述第一模型的输入信息对应的样本指示。

35 23、一种通信装置，包括：

通信单元，用于从第一设备或目标设备获取第二信息，其中，所述通信装置 700 上部署第一模型，所述第一设备或所述目标设备上存储有知识信息和知识信息的调用接口信息中的至少之一，其中，所述目标设备是使用所述调用接口信息所访问的设备，所述知识信息包括以下至少之一：知识图谱、知识向量库；

40 其中，所述第二信息根据第一信息和所述第一设备上存储或调用的知识信息确定；

其中，所述第一信息包括以下至少之一：

用于所述第一模型进行推理的辅助信息；

用于所述第一模型进行推理的需求信息；

所述第一模型的输入信息；

45 知识信息的调用接口信息；

所述第二信息的接收设备指示；

所述第一模型的输入信息对应的样本指示。

24、一种通信设备，包括：存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，所述计算机程序被所述处理器执行时实现如权利要求 1 至 8 中任一项所述的方法中的步骤，或如权利要求 9 至 11 中任一项所述的方法中的步骤，或者，
5 如权利要求 12 至 20 中任一项所述的方法中的步骤。

25、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 8 中任一项所述的方法中的步骤，或如权利要求 9 至 11 中任一项所述的方法中的步骤，或者，如权利要求 12 至 20 中任一项所述的方法中的步骤。
10

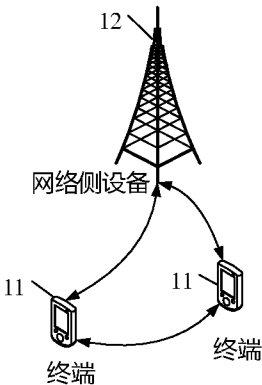


图 1

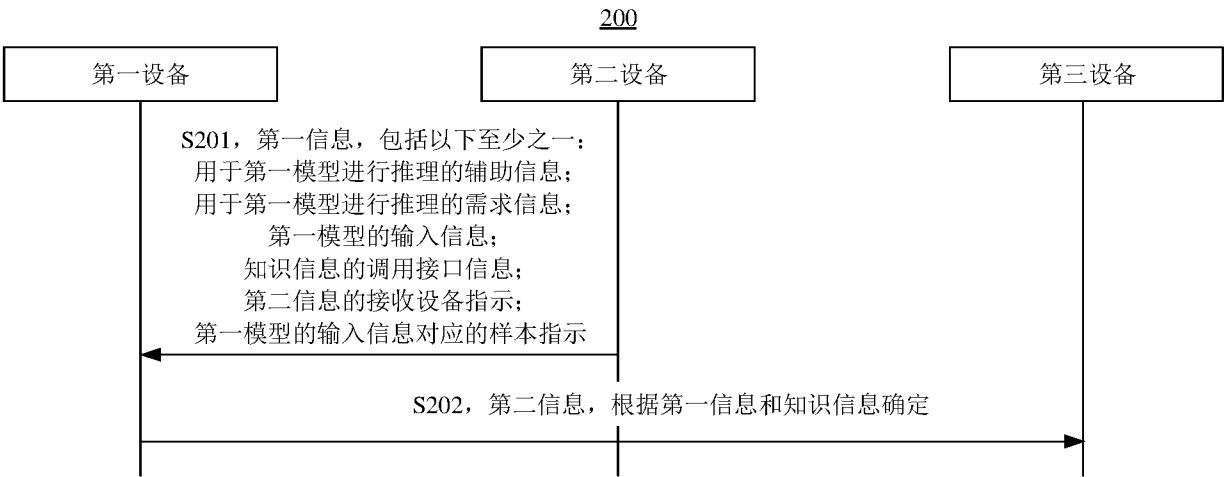


图 2

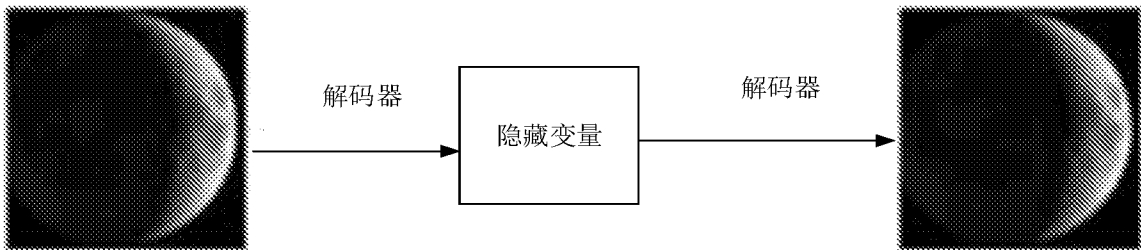


图 3

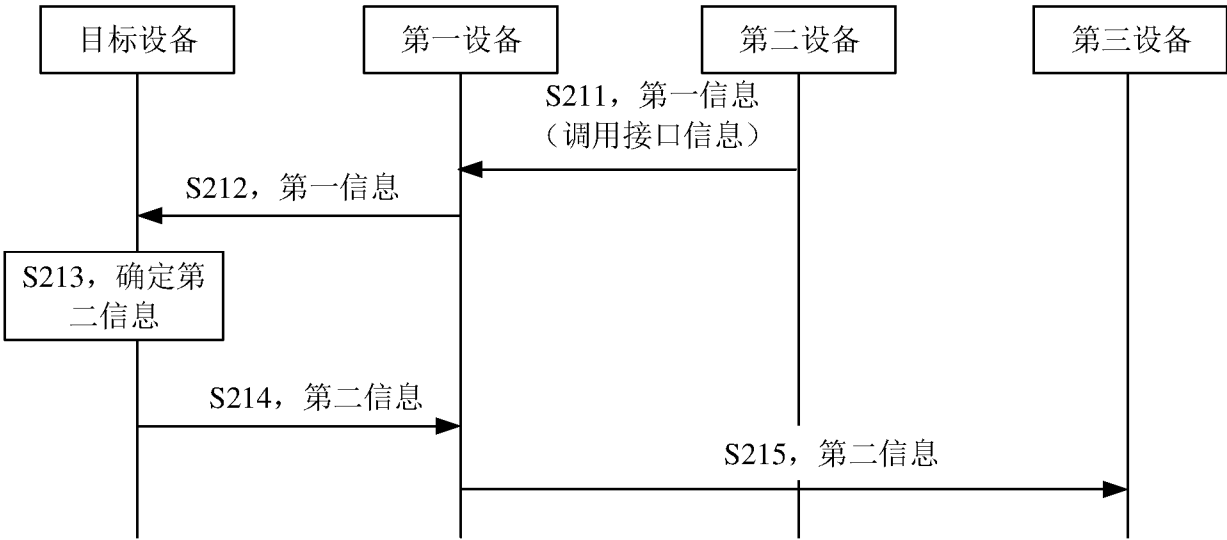


图 4

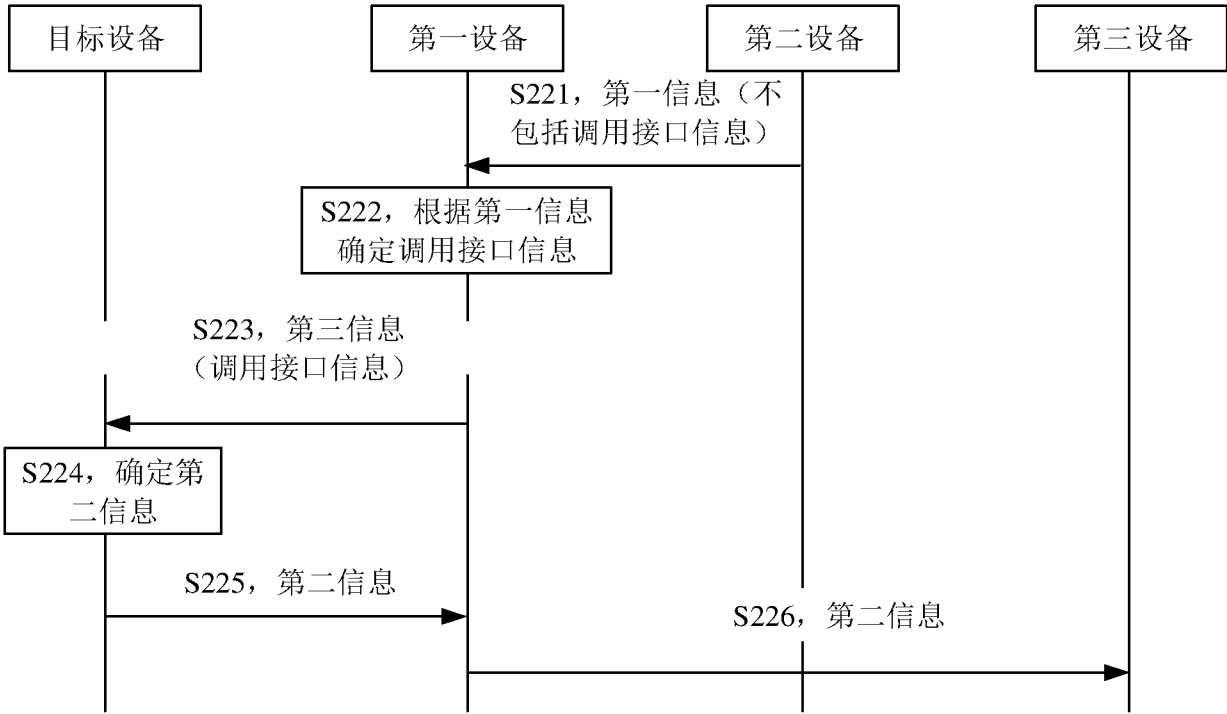


图 5

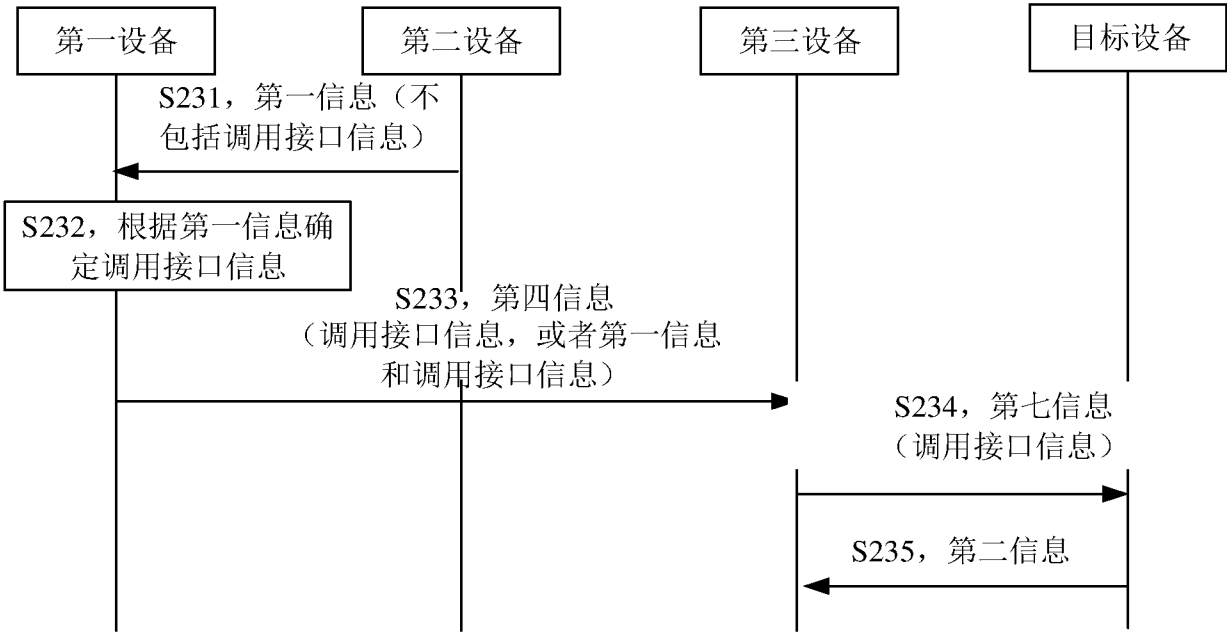


图 6

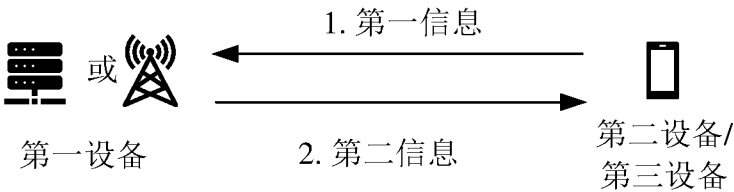


图 7

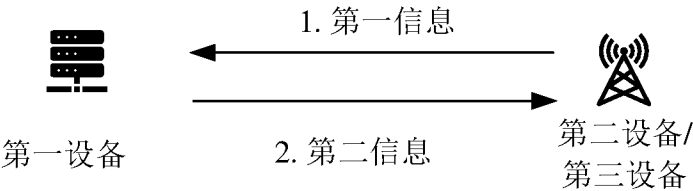


图 8

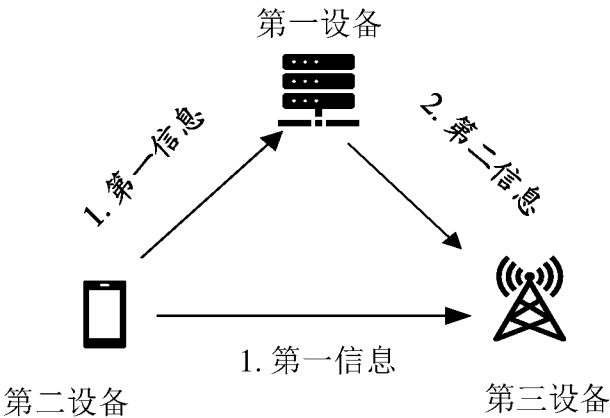


图 9

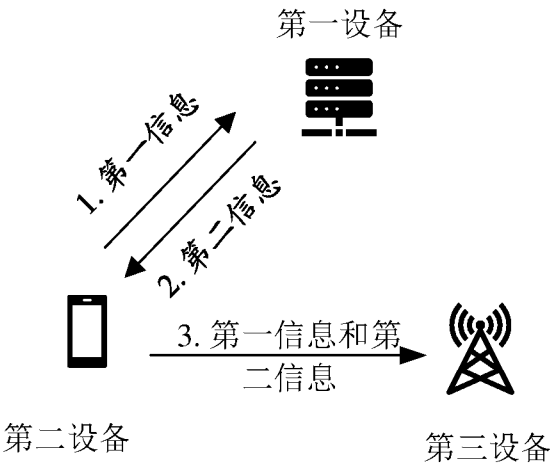


图 10

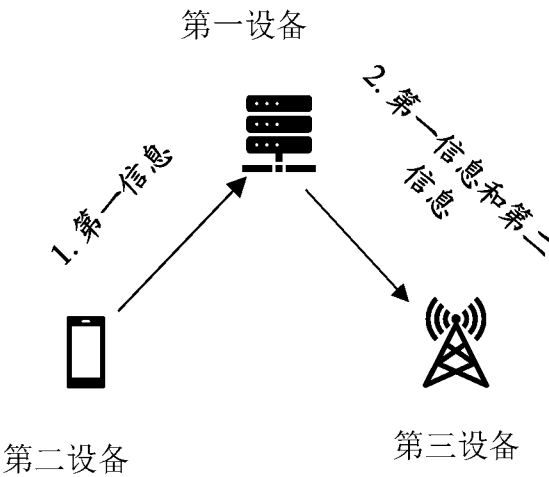


图 11

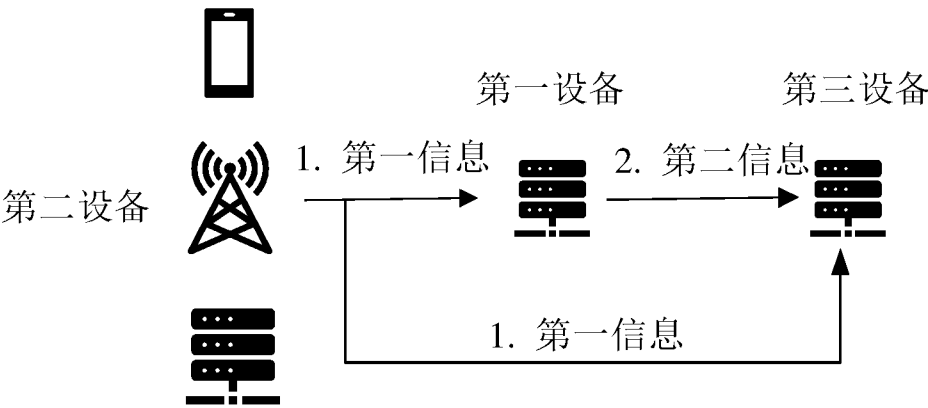


图 12

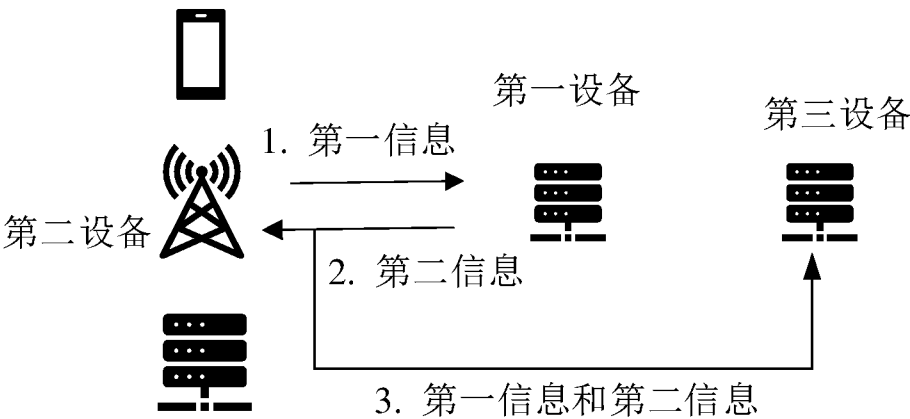


图 13

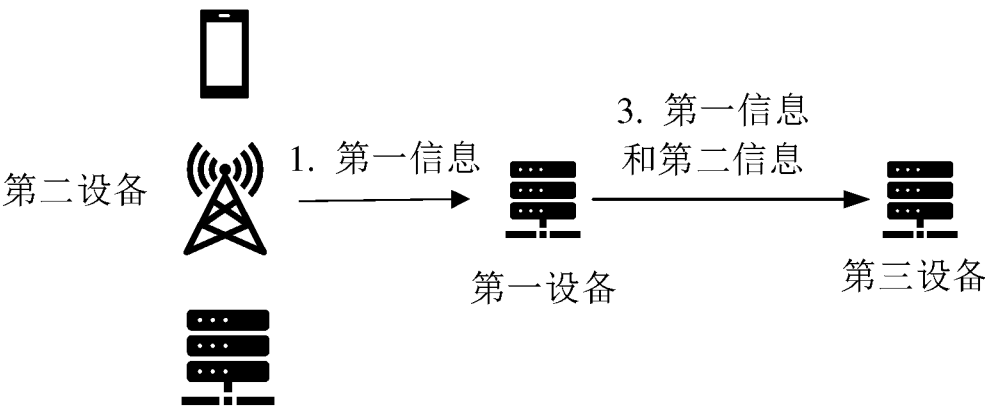


图 14



图 15



图 16



图 17

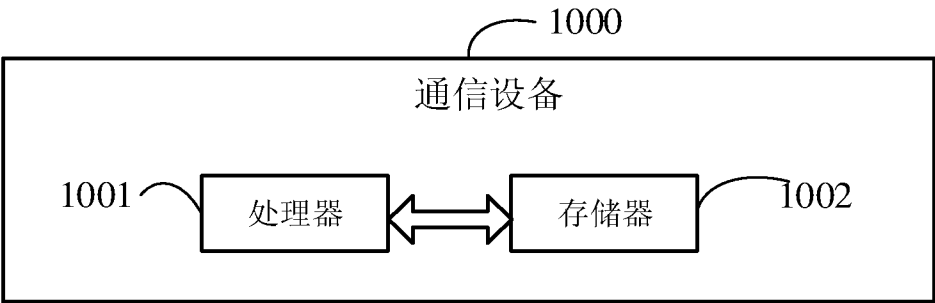


图 18

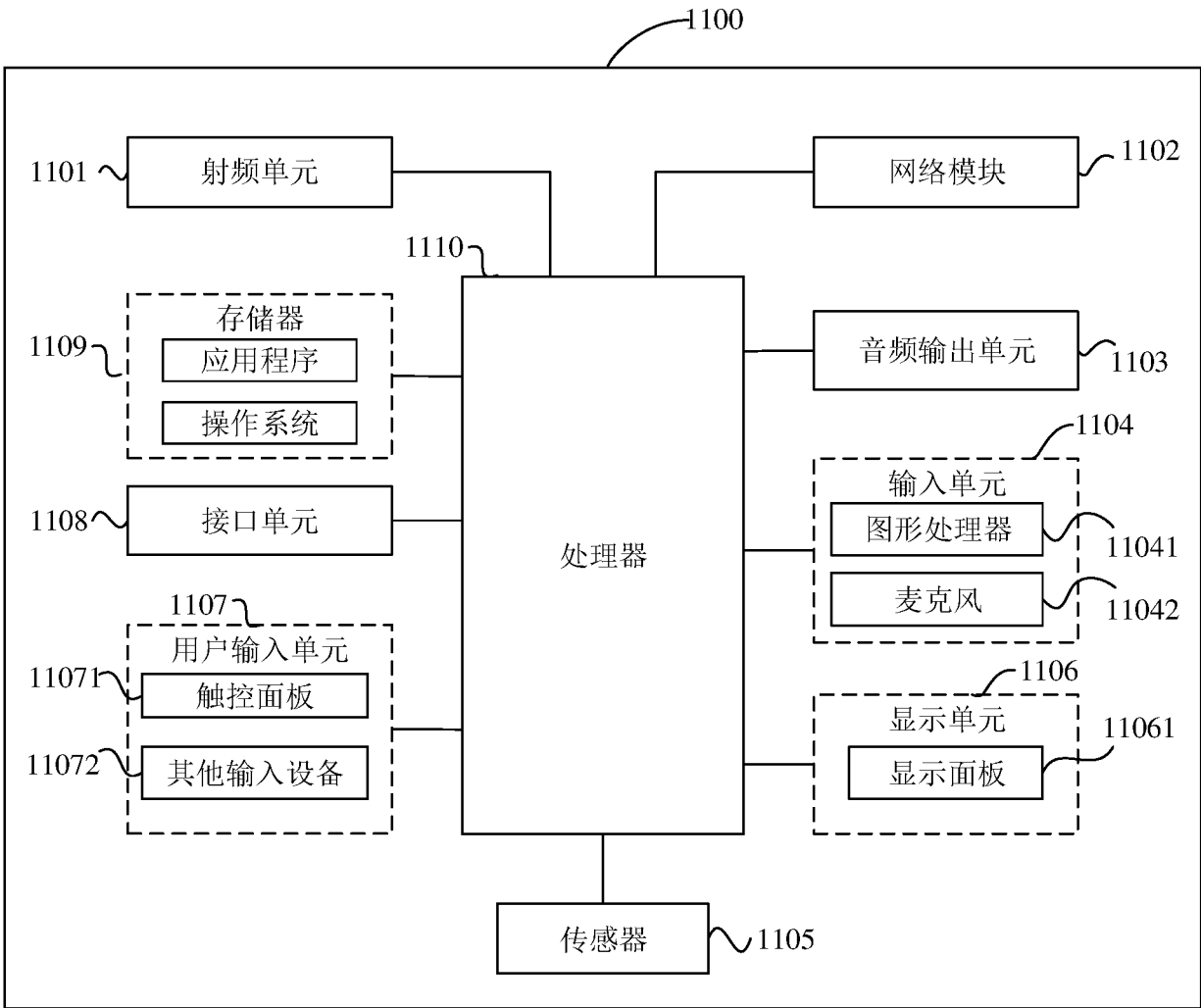


图 19

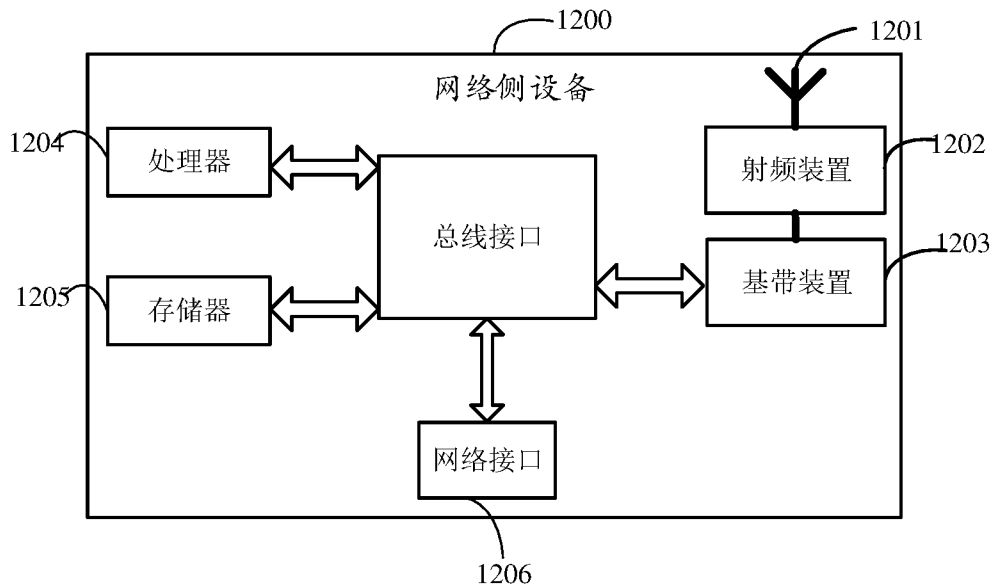


图 20

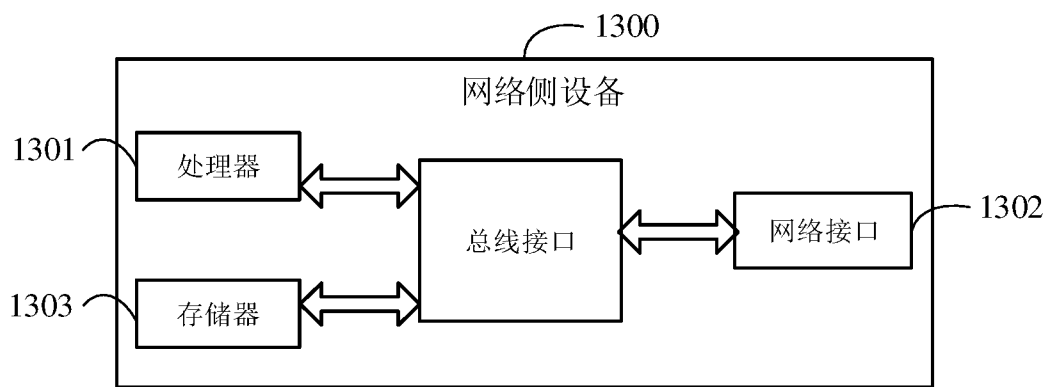


图 21

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2024/133006

A. CLASSIFICATION OF SUBJECT MATTER

G06N5/04(2023.01)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: G06N

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)

CNXTX, DWPL, ENTXT, ENTXTC, WPABS, WPABSC, CNKI: 维沃, 通信, 知识, 图谱, 人工智能, 模型, 深度学习, 机器学习, 神经网络, 发送, 接收, 接口, 推测, 推断, 推理, 预测; VIVO, communication, knowledge, graph, artificial intelligence, AI, model, deep learning, machine learning, neural network, send, transmit, receive, interface, API, inference, predict+

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 112805743 A (SAMSUNG ELECTRONICS CO., LTD.) 14 May 2021 (2021-05-14) description, paragraphs [0004], [0030]-[0140], figures 1-16	1-2, 5-13, 18-25
Y	CN 112287114 A (ZHUHAI DAHENGQIN TECHNOLOGY DEVELOPMENT CO., LTD.) 29 January 2021 (2021-01-29) description, paragraphs [0075]-[0142]	1-2, 5-13, 18-25
A	CN 111865446 A (CENTRAL SOUTH UNIVERSITY) 30 October 2020 (2020-10-30) entire document	1-25
A	CN 114205852 A (PURPLE MOUNTAIN LABORATORIES FOR NETWORK AND COMMUNICATION SECURITY) 18 March 2022 (2022-03-18) entire document	1-25
A	US 4999833 A (ITT CORP.) 12 March 1991 (1991-03-12) entire document	1-25
A	WO 2023186099 A1 (VIVO MOBILE COMMUNICATION CO., LTD.) 05 October 2023 (2023-10-05) entire document	1-25



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

07 February 2025

Date of mailing of the international search report

11 February 2025

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/CN2024/133006

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	112805743	A	14 May 2021	US	2020118010	A1	16 April 2020
				WO	2020080773	A1	23 April 2020
				EP	3821390	A1	19 May 2021
				EP	3821390	A4	25 August 2021
				KR	20200042739	A	24 April 2020
CN	112287114	A	29 January 2021	None			
CN	111865446	A	30 October 2020	CN	111865446	B	06 April 2021
CN	114205852	A	18 March 2022	CN	114205852	B	03 May 2022
				WO	2023155481	A1	24 August 2023
				EP	4482193	A1	25 December 2024
US	4999833	A	12 March 1991	None			
WO	2023186099	A1	05 October 2023	US	2025024330	A1	16 January 2025
				CN	116939713	A	24 October 2023

A. 主题的分类

G06N5/04(2023.01)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)

IPC: G06N

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,DWPI,ENTXT,ENTXTC,WPABS,WPABSC,CNKI:维沃, 通信, 知识, 图谱, 人工智能, 模型, 深度学习, 机器学习, 神经网络, 发送, 接收, 接口, 推测, 推断, 推理, 预测; VIVO, communication, knowledge, graph, artificial intelligence, AI, model, deep learning, machine learning, neural network, send, transmit, receive, interface, API, inference, predict+

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
Y	CN 112805743 A (三星电子株式会社) 2021年5月14日 (2021 - 05 - 14) 说明书第[0004], [0030]-[0140]段, 图1-16	1-2, 5-13, 18-25
Y	CN 112287114 A (珠海大横琴科技发展有限公司) 2021年1月29日 (2021 - 01 - 29) 说明书第[0075]-[0142]段	1-2, 5-13, 18-25
A	CN 111865446 A (中南大学) 2020年10月30日 (2020 - 10 - 30) 全文	1-25
A	CN 114205852 A (网络通信与安全紫金山实验室) 2022年3月18日 (2022 - 03 - 18) 全文	1-25
A	US 4999833 A (ITT) 1991年3月12日 (1991 - 03 - 12) 全文	1-25
A	WO 2023186099 A1 (维沃移动通信有限公司) 2023年10月5日 (2023 - 10 - 05) 全文	1-25

☐ 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

2025年2月7日

国际检索报告邮寄日期

2025年2月11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

王越

电话号码 (+86) 62089109

PCT/ISA/210 表(第2页) (2022年7月)

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)
CN	112805743	A	2021年5月14日	US	2020118010	A1	2020年4月16日
				WO	2020080773	A1	2020年4月23日
				EP	3821390	A1	2021年5月19日
				EP	3821390	A4	2021年8月25日
				KR	20200042739	A	2020年4月24日
CN	112287114	A	2021年1月29日	无			
CN	111865446	A	2020年10月30日	CN	111865446	B	2021年4月6日
CN	114205852	A	2022年3月18日	CN	114205852	B	2022年5月3日
				WO	2023155481	A1	2023年8月24日
				EP	4482193	A1	2024年12月25日
US	4999833	A	1991年3月12日	无			
WO	2023186099	A1	2023年10月5日	US	2025024330	A1	2025年1月16日
				CN	116939713	A	2023年10月24日