

(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
2023 年 6 月 29 日 (29.06.2023)



(10) 国际公布号
WO 2023/116259 A1

(51) 国际专利分类号:
G06N 20/00 (2019.01)

(21) 国际申请号: PCT/CN2022/131852

(22) 国际申请日: 2022 年 11 月 15 日 (15.11.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202111608193.5 2021 年 12 月 23 日 (23.12.2021) CN

(71) 申请人: 大唐移动通信设备有限公司 (**DATANG MOBILE COMMUNICATIONS EQUIPMENT CO., LTD.**) [CN/CN]; 中国北京市海淀区上地东路 5 号院 1 号楼 1 层, Beijing 100085 (CN)。

(72) 发明人: 刘莹莹 (**LIU, Yingying**); 中国北京市海淀区上地东路 5 号院 1 号楼 1 层, Beijing 100085 (CN)。段小嫣 (**DUAN, Xiaoyan**); 中国北京市海淀区上地东路 5 号院 1 号楼 1 层, Beijing 100085 (CN)。

(74) 代理人: 北京同立钧成知识产权代理有限公司 (**LEADER PATENT & TRADEMARK FIRM**); 中国北京市海淀区西直门北大街 32 号枫蓝国际 A 座 8F-6, Beijing 100082 (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, 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, 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,

(54) **Title:** MODEL SEGMENTATION ASSISTING METHOD AND APPARATUS, AND READABLE STORAGE MEDIUM

(54) 发明名称: 辅助模型切分的方法、装置及可读存储介质

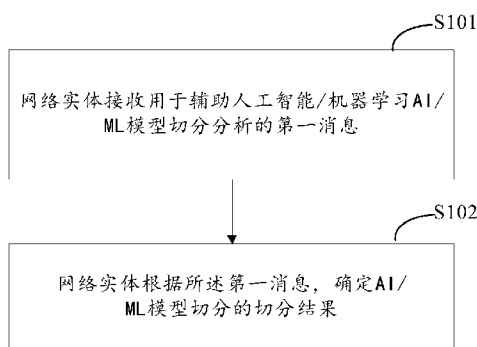


图 2

S101 A network entity receives a first message for assisting artificial intelligence/machine learning (AI/ML) model segmentation analysis
S102 The network entity determines a segmentation result of AI/ML model segmentation according to the first message

(57) **Abstract:** The present disclosure provides a model segmentation assisting method and apparatus, and a readable storage medium. The method comprises: a network entity receives a first message for assisting artificial intelligence/machine learning (AI/ML) model segmentation analysis, the first message comprising at least one of the following: a UE(s) that is to participate in AI/ML model segmentation or a subscription permanent identifier (SUPI), an identifier of an application that is to use an AI/ML model, computing power that can be provided by the UE(s) that is to participate in model joint reasoning, a memory that can be provided by the UE(s) that is to participate in the model joint reasoning, the remaining battery level of the UE(s) that is to participate in the model joint reasoning, and the size of the AI/ML model; and the network entity performs analysis according to the first message to determine a segmentation result of AI/ML model segmentation. According to the present disclosure, terminal capability-based model segmentation analysis can be achieved, and then protection of privacy of a terminal and an optimization method for network resources are effectively implemented.

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))。

(57) 摘要：本公开提供一种辅助模型切分的方法、装置及可读存储介质，该方法包括：网络实体接收用于辅助人工智能/机器学习AI/ML模型切分分析的第一消息，第一消息中包括下述至少一项：待参与AI/ML模型切分的UE(s)或者用户永久标识SUPI、待使用AI/ML模型的应用的标识、待参与模型联合推理的UE(s)可提供的算力、待参与模型联合推理的UE(s)可提供的内存、待参与模型联合推理的UE(s)剩余的电量、AI/ML模型的大小；网络实体根据第一消息进行分析，确定AI/ML模型切分的切分结果。本公开能够实现基于终端能力对模型切分进行分析，进而有效地实现对终端隐私的保护以及网络资源的优化本法。

辅助模型切分的方法、装置及可读存储介质

本公开要求于2021年12月23日提交中国专利局、申请号为202111608193.5、
申请名称为“辅助模型切分的方法、装置及可读存储介质”的中国专利申请
5 的优先权，其全部内容通过引用结合在本公开中。

技术领域

本公开涉及通信技术领域，尤其涉及一种辅助模型切分的方法、装置及可读存储介质。

背景技术

近年来，由于人工智能的技术突破，人工智能的应用越来越广泛。对于移动终端，由于其具有严格的能耗、计算和内存成本限制，无法在终端运行重量级的人工智能（英文为：Artificial Intelligence，简称为：AI）/机器学习（英文为：Machine Learning，简称为：ML）模型（以下称为AI/ML模型），
15 因此，目前采用的手段是将许多AI/ML模型的推理从移动终端传输到云或者其他终端，即需要把AI/ML模型传输给云或者其他终端。

此外，在SA#93e通过的SA1 R18需求中，需要AI/ML模型传输的场景如下：AI/ML端点之间的AI/ML模型切分，即一个AI/ML模型可以基于当前的任务或者环境切分成多个部分。趋势是将计算复杂，能耗大的部分由网络进行推理，需要隐私保护或者时延敏感的部分在终端推理。
20

但是，现有技术中，无法实现基于终端能力对模型切分的分析，进而无法有效地实现对终端隐私的保护以及网络资源的优化。

发明内容

本公开提供一种辅助模型切分的方法、装置及可读存储介质，解决了现有技术中无法实现基于终端能力对模型切分的分析，进而无法有效地实现对终端隐私的保护以及网络资源的优化的技术问题。
25

第一方面，本公开提供一种辅助模型切分的方法，应用于网络实体，所述方法包括：

接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息；

根据所述第一消息，确定 AI/ML 模型切分的切分结果。

- 5 可选的，所述第一消息中包括下述至少一项：待参与 AI/ML 模型切分的 UE(s)或者用户永久标识 SUPI、使用 AI/ML 模型的应用的标识、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量、AI/ML 模型的大小；

- 10 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第一消息中还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息。

可选地，所述根据所述第一消息，确定 AI/ML 模型切分的切分结果，包括：

- 15 根据所述第一消息中的待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量，确定所述待参与模型联合推理的 UE(s)对应的模型切分点信息；

- 20 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，则根据所述第一消息中的待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量、推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息，确定所述待参与模型联合推理的 UE(s)对应的模型切分点信息；

将所述模型切分点信息作为所述切分结果。

可选地，所述确定所述待参与模型联合推理的 UE(s)对应的模型切分点信息，包括：

- 25 将所有确定参与执行模型联合推理的 UE(s)中的任一 UE 作为目标 UE，针对每个所述目标 UE 执行下述步骤：

若所述目标 UE 可提供的算力低于第一预设算力阈值或所述目标 UE 可提供的内存低于第一预设内存阈值或所述目标 UE 剩余的电量低于第一预设电量阈值时，确定所述目标 UE 执行第一预设数目层的推理；

- 30 若所述目标 UE 可提供的算力高于第一预设算力阈值并低于第二预设算

力阈值，和所述目标 UE 可提供的内存高于第一预设内存阈值并低于第二预设内存阈值，以及所述目标 UE 剩余的电量高于第一预设电量阈值并低于第二预设电量阈值时，确定所述目标 UE 执行第二预设数目的推理，第一预设数目层小于第二预设数目层；

- 5 若所述目标 UE 可提供的算力高于第二预设算力阈值并低于第三预设算力阈值，和所述目标 UE 可提供的内存高于第二预设内存阈值并低于第三预设内存阈值，以及所述目标 UE 剩余的电量高于第二预设电量阈值并低于第三预设电量阈值时，确定所述目标 UE 执行第三预设数目层的推理，第二预设数目层小于第三预设数目层；以此类推，直至确定所述目标 UE 执行第 N+1
10 预设数目层的推理，第 N 预设数目层小于第 N+1 预设数目层；

 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，且若推理 AI/ML 模型需要的数据集的隐私等级低于第一预设隐私等级或模型不同层的时延需求信息低于第一预设时延，则所述目标 UE 执行所述第一预设数目层的推理；

- 15 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，且若推理 AI/ML 模型需要的数据集的隐私等级高于第一预设隐私等级并低于第二预设隐私等级，和模型不同层的时延需求信息高于第一预设时延并低于第二预设时延时，确定所述目标 UE 执行第二预设数目层的推理；

- 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，且
20 若推理 AI/ML 模型需要的数据集的隐私等级高于第二预设隐私等级并低于第三预设隐私等级，和模型不同层的时延需求信息高于第二预设时延并低于第三预设时延时，确定所述目标 UE 执行第三预设数目层的推理；以此类推，直至确定所述目标 UE 执行第 N+1 预设数目层的推理，第 N 预设数目层小于第 N+1 预设数目层；

- 25 根据预设数目层确定 AI/ML 模型切分点信息，所述 AI/ML 模型切分点信息用于表示 AI/ML 模型切分比例，所述预设数目层包括第 N 预设数目层，N 大于或等于一。

- 本公开实施例中，网络实体通过接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息，并根据第一消息中的待参与模型联合推理的 UE
30 可提供的算力、待参与模型联合推理的 UE 可提供的内存、待参与模型联合

推理的 UE 剩余的电量来确定切分 AI/ML 模型的模型切分点信息，如果推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延（其中，如果 UE 无法提供模型不同层的时延需求信息，则模型不同层的时延需求信息由 AF 提供；如果 UE 无法提供需要的数据集的隐私等级，基于 SA3 确认是否可以设定数据集的隐私等级，比如，如果 SA3 可以提供隐私等级设置，则可以由 SA3 提供数据集的隐私等级），还可以结合推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息来确定切分 AI/ML 模型的模型切分点信息，作为切分结果，用以实现网络实体或待参与模型联合推理的 UE 基于切分结果对模型进行联合推理操作，因此，实现了基于终端（即 UE）能力对模型切分的分析，进而有效地实现对终端隐私的保护以及网络资源的优化。

可选地，如果所述网络实体是网络数据分析功能 NWDAF，所述接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息，包括：

直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第一请求，所述第一请求用于请求对 AI/ML 模型切分进行分析；其中，所述第一请求中携带的参数包括下述至少一项：与模型标识或模型切分标识关联的分析类型标识、接受 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足切分条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量；若推理 AI/ML 模型需要数据集的隐私等级和/或模型不同层的时延，所述第一请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

根据所述第一请求中携带的参数，向 5GC NF(s)发送第二消息，所述第二消息用于请求 5GC NF(s)采集 UE(s)对应的第一数据，所述第一数据中包括下述至少一项：待参与 AI/ML 模型切分的 UE(s)以及 SUPI、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量、AI/ML 模型的大小；若第一请求中携带的参数包括推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息，则所述第一数据中还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

接收 5GC NF(s)发送的所述第一数据，并将所述第一数据作为所述第一消息；所述第一数据是由待参与模型联合推理的 UE(s)同意所述第二消息的请求后提供的。

可选地，所述确定 AI/ML 模型切分的切分结果之后，所述方法还包括：

5 向所述 AF 发送第三消息，所述第三消息中包括 AI/ML 模型切分点信息以及所述第一数据；所述第三消息用于为 AF 提供向待参与模型联合推理的 UE(s)发送第二请求时携带的参数，所述第二请求用于请求与待参与模型联合推理的 UE(s)执行模型联合推理操作，所述第二请求中携带的参数包括所述第三消息。

10 本公开实施例中，当 UE(s)和 AF(s)联合推理模型(s)时，AF 请求 NWDAF 收集 UE 的能力进行模型切分点的判断，NWDAF 收集并判断后将结果分别反馈给 AF，AF 发给相关的 UE。

可选地，如果所述网络实体是网络数据分析功能 NWDAF，所述接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息，包括：

15 直接或通过网络能力开放功能 NEF 接收 AF 发送的第三请求；所述第三请求是由 AF 确定与待参与模型联合推理的 UE(s)具备完全支持执行模型联合推理的能力后确定的，其中，是否具备完全支持执行模型联合推理的能力是由 AF 根据接收到的待参与模型联合推理的 UE(s)发送的第四请求中携带的参数确定的，所述第四请求用于待参与模型联合推理的 UE(s)请求与 AF 进行模型联合推理，所述第四请求中携带的参数包括下述至少一项：与模型标识或模型切分标识关联的分析类型标识、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第四请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的
20 隐私等级和/或模型不同层的时延需求信息，所述第三请求中携带的参数包括所述第四请求中携带的参数；

将所述第三请求中携带的参数作为所述第一消息；

其中，若第四请求中携带的参数中不包含模型不同层的时延需求信息，则模型不同层的时延需求信息由 AF 提供。

30 可选地，所述确定 AI/ML 模型切分的切分结果之后，所述方法还包括：

向所述 AF 发送第四消息，所述第四消息中包括所述 AI/ML 模型切分点信息、所述第一消息中的数据；所述第四消息用于为 AF 提供向待参与模型联合推理的 UE(s)发送的第四请求结果，所述第四请求结果中包括所述 AI/ML 模型切分点信息。

- 5 本公开实施例中，当 UE(s)和 AF(s)联合推理（多个）模型时，UE 上报自身的能力给 AF，AF 请求 NWDAF 进行模型切分点的判断，NWDAF 把判断结果反馈给 AF，AF 发送给 UE。

可选地，如果所述网络实体是网络数据分析功能 NWDAF，所述接收用于请求人工智能/机器学习 AI/ML 模型切分分析的第一消息，包括：

- 10 直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第五请求，所述第五请求是由 AF 确定与待参与模型联合推理的 UE(s)不具备完全支持执行模型联合推理的能力后确定的，其中，是否具备完全支持执行模型联合推理的能力是由 AF 根据接收到的待参与模型联合推理的 UE(s)发送的第六请求确定的，所述第五请求用于请求对 AI/ML 模型切分进行分析以及请求查找其他可参与模型推理的 UE(s)并提供其他可参与模型推理的 UE(s)可提供的算力、
15 可提供的内存、剩余的电量；其中，所述第五请求中携带的参数包括下述至少一项：与模型切分关联的分析类型标识、接受 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足切分条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE(s)可提供的算力、待
20 参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第五请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；所述第六请求用于待参与模型联合推理的 UE(s)请求与 AF 进行模型联合推理，所述第六请求中携带的参数包括所述
25 第五请求中携带的参数；

- 根据所述第五请求中携带的参数，向 5GC NF(s)发送第五消息，所述第五消息用于请求 5GC NF(s)采集 UE(s)对应的第二数据，所述第二数据中包括下述至少一项：待参与 AI/ML 模型切分的 UE(s)或者 SUPI、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待
30 参与模型联合推理的 UE(s)剩余的电量、AI/ML 模型的大小；若第五请求中

携带的参数包括推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息，则所述第二数据中还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

接收 5GC NF(s)发送的所述第二数据，并将所述第二数据作为所述第一消息；所述第二数据是由待参与模型联合推理的 UE(s)同意所述第五消息的请求后提供的；

其中，若第六请求中携带的参数中不包含模型不同层的时延需求信息，则模型不同层的时延需求信息由 AF 提供。

可选地，所述确定 AI/ML 模型切分的切分结果之后，所述方法还包括：
向所述 AF 发送第六消息，所述第六消息中至少包括所述 AI/ML 模型切分点信息以及所述第二数据；所述第六消息用于为 AF 提供向待参与模型联合推理的 UE(s)发送的第六请求结果，所述第六请求结果中包括所述待参与模型联合推理的 UE(s)以及对应的所述 AI/ML 模型切分点信息。

本公开实施例中，当 UE(s)和 AF(s)联合推理模型时，UE 上报自身的能力给 AF，AF 请求 NWDAF 进行模型切分点的判断并请求其他可参与模型推理的 UE(s)参与本模型的推理，NWDAF 把收集和判断结果反馈给 AF，AF 发送给 UE。

可选地，如果所述网络实体是网络数据分析功能 NWDAF，所述接收用于请求人工智能/机器学习 AI/ML 模型切分分析的第一消息，包括：

接收待参与模型联合推理的 UE(s)发送的第七请求，所述第七请求用于请求与 NWDAF 执行模型联合推理操作；所述第七请求中携带的参数包括下述至少一项：待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第七请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；根据所述第七请求中携带的参数，确定 AF 与所述待参与模型联合推理的 UE(s)是否具备完全支持执行模型联合推理的能力；

若与所述待参与模型联合推理的 UE(s)具备完全支持执行模型联合推理的能力，则将所述第七请求中携带的参数作为所述第一消息；

若与所述待参与模型联合推理的 UE(s)不具备完全支持执行模型联合推

理的能力,则请求 NF 查找其他可参与模型推理的 UE(s)并提供其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量,且若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,则请求提供其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

接收 NF 发送的查找结果,所述查找结果包括其他可参与模型推理的 UE(s)以及其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量,若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,则查找结果还包括:其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

将所述第七请求中携带的参数以及所述查找结果作为所述第一消息;

其中,若第七请求中携带的参数中不包含模型不同层的时延需求信息,则模型不同层的时延需求信息由 AF 提供。

可选地,所述确定 AI/ML 模型切分的切分结果之后,所述方法还包括:向所述待参与模型联合推理的 UE(s)发送第七请求结果,所述第七请求结果中包括 AI/ML 模型切分点信息;

其中,若存在其他参与联合模型推理的 UE(s),则所述第七请求结果中还包括其他参与联合模型推理的 UE(s)对应的模型切分点信息,并通过 AF 透传给所述 UE(s)。

本公开实施例中,当 UE 和 NWDAF 联合推理模型时,UE 通过 AF 上报自身的能力 NWDAF,同时请求 NWDAF 进行模型切分点的判断,NWDAF 把判断结果通过 AF 反馈给 UE。

可选地,如果所述网络实体是应用功能 AF,所述接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息,包括:

接收待参与模型联合推理的 UE(s)发送的第八请求,所述第八请求用于请求与 AF 执行模型联合推理操作;所述第八请求中携带的参数包括下述至少一项:与模型标识或模型切分标识关联的分析类型标识、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量;若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,

所述第八请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；根据所述第八请求中携带的参数，确定与所述待参与模型联合推理的 UE(s)是否具备完全支持执行模型联合推理的能力；

5 若与所述待参与模型联合推理的 UE(s)具备完全支持执行模型联合推理的能力，则将所述第八请求中携带的参数作为所述第一消息；

若与所述待参与模型联合推理的 UE(s)不具备完全支持执行模型联合推理的能力，则向网络数据分析功能 NWDAF 发送第九请求，所述第九请求中携带的参数包括所述第八请求中携带的参数，且所述第九请求用于请求
10 NWDAF 通过 NF 查找其他可参与模型推理的 UE(s)以及其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量，若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第九请求中携带的参数还包括：其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

15 接收 NWDAF 发送的第九请求结果，并将所述第九请求结果以及所述第八请求中携带的参数作为所述第一消息；其中，所述第九请求结果包括其他可参与模型推理的 UE(s)以及其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量，若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第九请求结果还包括：其他可参与模型推理的 UE(s)
20 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

若第八请求中携带的参数中不包含模型不同层的时延需求信息，则模型不同层的时延需求信息由 AF 提供。

可选地，所述确定 AI/ML 模型切分的切分结果之后，所述方法还包括：

25 将确定与所述待参与模型联合推理的 UE(s)是否具备完全支持执行模型联合推理的能力的结果、所述待参与模型联合推理的 UE(s)的相关信息以及所述待参与模型联合推理的 UE(s)对应的模型切分点信息发送给所述 UE；

其中，所述相关信息包括下述至少一项：待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或
30 模型不同层的时延，所述相关信息中还包括：推理 AI/ML 模型需要的数据集

的隐私等级和/或模型不同层的时延需求信息。

本公开实施例中，当 UE 和 AF 联合推理模型时，UE 上报自身的能力给 AF，AF 进行联合推理和模型切分点的判断，然后把判断结果反馈给 UE。

可选地，如果所述网络实体是应用功能 AF，所述接收用于辅助人工智能
5 /机器学习 AI/ML 模型切分分析的第一消息，包括：

向网络数据分析功能 NWDAF 发送第十请求，所述第十请求用于请求
NWDAF 向 5GC NF(s)采集 UE(s)对应的用于分析 AI/ML 模型切分的数据，所
述第十请求中携带的参数包括下述至少一项：与模型标识或模型切分标识关
10 联的分析类型标识、接收 AI/ML 模型切分的一个用户设备 UE 或一组 UE 的
标识或满足分析条件的任意 UE 的、AI/ML 模型切分的区域、AI/ML 模型的大
小、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)
可提供的内存、待参与模型联合推理的 UE(s)剩余的电量；若推理 AI/ML 模
型时需要数据集的隐私等级和/或模型不同层的时延，所述第十请求中携带的
15 参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时
延需求信息；

接收 NWDAF 发送的第十请求结果，所述第十请求结果包括下述至少一
项：待参与 AI/ML 模型切分的 UE(s)或者 SUPI、使用 AI/ML 模型的应用的
标识、待参与模型联合推理的 UE 可提供的算力、待参与模型联合推理的 UE(s)
可提供的内存、待参与模型联合推理的 UE(s)剩余的电量、AI/ML 模型的大
20 小；若第十请求中携带的参数包括推理 AI/ML 模型需要的数据集的隐私等级
和/或模型不同层的时延需求信息，则所述第十请求结果中还包括：推理
AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

将所述第十请求结果作为所述第一消息。

可选地，所述确定 AI/ML 模型切分的切分结果之后，所述方法还包括：
25 向待参与模型联合推理的 UE(s)发送第十一请求，所述第十一请求用于请
求与所述待参与模型联合推理的 UE(s)执行模型联合推理操作；其中，所述第
十一请求中携带的参数包括下述至少一项：模型切分点信息、所述第一消息；

接收待参与模型联合推理的 UE(s)发送的第十一请求结果，所述第十一请
求结果是由待参与模型联合推理的 UE(s)根据第十一请求中携带的参数确定
30 的，所述第十一请求结果包括接受第十一请求或不接受第十一请求。

本公开实施例中，当 UE 和 AF 联合推理模型时，AF 请求 NWDAF 收集 UE 的能力，AF 基于收到的分析结果，进行模型切分点的判断，并向相关 UE 发起联合推理请求。

可选地，如果所述网络实体是新的网络实体(例如，MMF，Model Management Function)，所述接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息，包括：

直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第十二请求，所述第十二请求用于请求 AI/ML 模型切分分析；所述第十二请求中携带的参数包括下述至少一项：与模型标识或模型切分标识关联的分析类型标识、接收 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足分析条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第十二请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

根据所述第十二请求中携带的参数，向 NWDAF 发送第十三请求，所述第十三请求中携带的参数包括所述第十二请求中携带的参数，且所述第十三请求用于请求 NWDAF 向 5GC NF(s)采集 UE(s)的用于分析 AI/ML 模型切分的数据；

接收 NWDAF 发送的第十三请求结果；其中，所述第十三请求结果包括下述至少一项：待参与 AI/ML 模型切分的 UE(s)或者 SUPI、使用 AI/ML 模型的应用的标识、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量、AI/ML 模型的大小；若第十三请求中携带的参数包括推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息，则所述第十三请求结果中还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

将所述第十三请求结果作为所述第一消息。

可选地，所述确定 AI/ML 模型切分的切分结果之后，所述方法还包括：

向所述 AF 发送第八消息，所述第八消息中包括所述 AI/ML 模型切分点、

所述第一消息中的数据；所述第八消息用于为 AF 提供向待参与模型联合推理的 UE(s)发送第十四请求时携带的参数，所述第十四请求用于请求与待参与模型联合推理的 UE(s)执行模型联合推理操作，所述第十四请求中携带的参数包括所述第八消息。

- 5 本公开实施例中，当 UE 和 AF 联合推理模型时，AF 请求新的网络实体 MMF 进行模型切分，MMF 请求 NWDAF 收集 UE 的能力，MMF 基于收到的分析结果，进行模型切分点的判断，并向 AF 反馈结果，AF 向相关 UE 发起联合推理请求。

10 可选地，如果所述网络实体是新的网络实体 MMF，所述接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息，包括：

15 直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第十五请求，所述第十五请求用于请求 AI/ML 模型切分分析，所述第十五请求中携带的参数包括下述至少一项：与模型标识或模型切分标识关联的分析类型标识、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第十五请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

20 根据所述第十五请求中携带的参数，确定 AF 与所述待参与模型联合推理的 UE(s)是否具备完全支持执行模型联合推理的能力；

 若与所述待参与模型联合推理的 UE(s)具备完全支持执行模型联合推理的能力，则将所述第十五请求中携带的参数作为所述第一消息；

25 若与所述待参与模型联合推理的 UE(s)不具备完全支持执行模型联合推理的能力，则向 NWDAF 发送第十六请求，所述第十六请求中携带的参数包括所述第十五请求中携带的参数，且所述第十六请求用于请求 NWDAF 通过 NF 查找其他可参与模型推理的 UE(s)以及其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量，若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第十六请求中携带的参数还包括：其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

30

接收 NWDAF 发送的第十六请求结果，并将所述第十六请求结果以及所述第十五请求中携带的参数作为所述第一消息；其中，所述第十六请求结果包括其他可参与模型推理的 UE(s)以及其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量，若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第十六请求结果还包括：其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息。

可选地，所述确定 AI/ML 模型切分的切分结果之后，所述方法还包括：

向所述 AF 发送第十五请求结果，所述第十五请求结果中包括 AI/ML 模型切分点信息；

其中，若存在其他参与联合模型推理的 UE(s)，则所述第十五请求结果中还包括其他参与联合模型推理的 UE(s)对应的模型切分点信息，并通过 AF 透传给所述待参与模型联合推理的 UE(s)。

本公开实施例中，当 UE 和 AF 联合推理模型时，UE 上报自身能力给 AF 请求模型联合推理，AF 将自身以及 UE 能力传给 MMF，请求 MMF 进行模型切分，MMF 判断如果需要其他 UE(s)参与模型联合推理，则 MMF 请求 NWDAF 收集其他 UE(s)的信息，MMF 基于收集的信息进行模型切分并将结果反馈给 AF。

可选地，如果所述网络实体是策略控制功能 PCF，所述接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息，包括：

直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第十七请求，所述第十七请求用于请求 AI/ML 模型切分分析；所述第十七请求中携带的参数包括下述至少一项：与模型标识或模型切分标识关联的分析类型标识、接收 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足分析条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第十七请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

根据所述第十七请求中携带的参数，向网络数据分析功能 NWDAF 发送

第十八请求, 所述第十八请求中携带的参数包括所述第十七请求中携带的参数, 且所述第十八请求用于请求 NWDAF 向 5GC NF(s)采集 UE(s)对应的第五数据; 所述第五数据中包括下述至少一项: 待参与 AI/ML 模型切分的 UE(s)或者 SUPI、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量、AI/ML 模型的大小; 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 所述第五数据中还包括: 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

接收 NWDAF 发送的第十八请求结果; 其中, 所述第十八请求结果包括所述第五数据;

将所述第十八请求结果作为所述第一消息。

可选地, 所述确定 AI/ML 模型切分的切分结果之后, 所述方法还包括: 向所述 AF 发送第九消息, 所述第九消息中包括所述 AI/ML 模型切分点、所述第一消息的数据; 所述第九消息用于为 AF 提供向待参与模型联合推理的 UE(s)发送第十九请求时携带的参数, 所述第十九请求用于请求与待参与模型联合推理的 UE(s)执行模型联合推理操作, 所述第十九请求中携带的参数包括所述第九消息。

本公开实施例中, 当 UE 和 AF 联合推理模型时, AF 请求 PCF 进行模型切分的策略判断, PCF 请求 NWDAF 收集 UE 的能力, PCF 基于收到的分析结果, 进行模型切分点的判断, 并向 AF 反馈结果, AF 向相关 UE 发起联合推理请求。

可选地, 如果所述网络实体是策略控制功能 PCF, 所述接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息, 包括:

直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第二十请求, 所述第二十请求用于请求 AI/ML 模型切分分析, 所述第二十请求中携带的参数包括下述至少一项: 与模型标识或模型切分标识关联的分析类型标识、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量; 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 所述第二十请求中携带的参数还包括: 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延

需求信息;

根据所述第二十请求中携带的参数,确定所述 AF 以及待参与模型联合推理的 UE(s)是否具备完全支持执行模型联合推理的能力;

若所述 AF 以及待参与模型联合推理的 UE(s)具备完全支持执行模型联合推理的能力,则将所述第二十请求中携带的参数作为所述第一消息;

若所述 AF 以及待参与模型联合推理的 UE(s)不具备完全支持执行模型联合推理的能力,则向 NWDAF 发送第二十一请求,所述第二十一请求中携带的参数包括所述第二十请求中携带的参数,且所述第二十一请求用于请求 NWDAF 通过 NF 查找其他可参与模型推理的 UE(s)以及其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量,若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,所述第二十一请求中携带的参数还包括:其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

接收 NWDAF 发送的第二十一请求结果,并将所述第二十一请求结果以及所述第二十请求中携带的参数作为所述第一消息;其中,所述第二十一请求结果包括:其他可参与模型推理的 UE(s)以及其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量,若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,所述第二十一请求结果还包括:其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息

若第二十请求中携带的参数中不包含模型不同层的时延需求信息,则模型不同层的时延需求信息由 AF 提供。

可选地,所述确定 AI/ML 模型切分的切分结果之后,所述方法还包括:

向所述 AF 发送第二十请求结果,所述第二十请求结果中包括 AI/ML 模型切分点信息;

其中,若存在其他参与联合模型推理的 UE(s),则所述第二十请求结果中还包括其他参与联合模型推理的 UE(s)对应的模型切分点信息,并通过 AF 透传给所述 UE(s)。

本公开实施例中,当 UE 和 AF 联合推理模型时,UE 上报自身能力给 AF 请求模型联合推理,AF 将自身以及 UE 能力发给 PCF,请求 PCF 进行模型切

分, PCF 判断如果需要其他 UE(s)参与模型联合推理, 则 PCF 请求 NWDAF 收集其他 UE(s)的信息, PCF 基于收集的信息进行模型切分并将结果反馈给 AF。

第二方面, 本公开提供一种辅助模型切分的方法, 所述方法应用于用户设备 UE, 所述方法包括:

根据自身能力信息, 确定 AI/ML 模型切分点信息;

将所述 AI/ML 模型切分点信息作为 AI/ML 模型切分的切分结果。

可选地, 所述根据自身能力信息, 确定 AI/ML 模型切分点信息, 包括:

根据自身能力信息中的可提供的算力、可提供的内存以及剩余的电量,

10 确定模型切分点信息;

若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 则根据所述自身能力信息中的可提供的算力、可提供的内存、剩余的电量、推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息, 确定模型切分点信息。

15 可选地, 所述确定模型切分点信息, 包括:

将所有确定参与执行模型联合推理的 UE(s)中的任一 UE 作为目标 UE, 针对每个所述目标 UE 执行下述步骤:

若所述目标 UE 可提供的算力低于第一预设算力阈值或所述目标 UE 可提供的内存低于第一预设内存阈值或所述目标 UE 剩余的电量低于第一预设电量阈值时, 确定所述目标 UE 执行第一预设数目层的推理;

25 若所述目标 UE 可提供的算力高于第一预设算力阈值并低于第二预设算力阈值, 和所述目标 UE 可提供的内存高于第一预设内存阈值并低于第二预设内存阈值, 以及所述目标 UE 剩余的电量高于第一预设电量阈值并低于第二预设电量阈值时, 确定所述目标 UE 执行第二预设数目层的推理, 第一预设数目层小于第二预设数目层;

30 若所述目标 UE 可提供的算力高于第二预设算力阈值并低于第三预设算力阈值, 和所述目标 UE 可提供的内存高于第二预设内存阈值并低于第三预设内存阈值, 以及所述目标 UE 剩余的电量高于第二预设电量阈值并低于第三预设电量阈值时, 确定所述目标 UE 执行第三预设数目层的推理, 第二预设数目层小于第三预设数目层; 以此类推, 直至确定所述目标 UE 执行第 N+1

预设数目层的推理，第 N 预设数目层小于第 N+1 预设数目层；

若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，且若推理 AI/ML 模型需要的数据集的隐私等级低于第一预设隐私等级或模型不同层的时延需求信息低于第一预设时延，则所述目标 UE 执行所述第一预设数目层的推理；

若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，且若推理 AI/ML 模型需要的数据集的隐私等级高于第一预设隐私等级并低于第二预设隐私等级，和模型不同层的时延需求信息高于第一预设时延并低于第二预设时延时，确定所述目标 UE 执行第二预设数目层的推理；

若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，且若推理 AI/ML 模型需要的数据集的隐私等级高于第二预设隐私等级并低于第三预设隐私等级，和模型不同层的时延需求信息高于第二预设时延并低于第三预设时延时，确定所述目标 UE 执行第三预设数目层的推理；以此类推，直至确定所述目标 UE 执行第 N+1 预设数目层的推理，第 N 预设数目层小于第 N+1 预设数目层；

根据预设数目层确定 AI/ML 模型切分点信息，所述 AI/ML 模型切分点信息用于表示 AI/ML 模型切分比例，所述预设数目层包括第 N 预设数目层，N 大于或等于一。

本公开实施例中，根据自身能力信息中的可提供的算力、可提供的内存以及剩余的电量，确定模型切分点信息；如果推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，还可以结合自身能力信息中的推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息来确定切分 AI/ML 模型的模型切分点信息，作为切分结果，用以实现网络实体或待参与模型联合推理的 UE 基于切分结果对模型进行联合推理操作，因此，实现了基于终端（即 UE）能力对模型切分的分析，进而有效地实现对终端隐私的保护以及网络资源的优化。

可选地，所述确定 AI/ML 模型切分的切分结果之后，所述方法还包括：

向应用功能 AF 发送第一请求，所述第一请求用于请求与所述 AF 执行模型联合推理操作；其中，所述第一请求中携带的参数包括下述至少一项 AI/ML 模型切分点、与模型标识或模型切分标识关联的分析类型标识、接收 AI/ML

模型切分的一个用户设备 UE 或一组 UEs 的标识或满足分析条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE 可提供的算力、待参与模型联合推理的 UE 可提供的内存、待参与模型联合推理的 UE 剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第一请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

接收 AF 发送的第一请求结果，所述第一请求结果是由 AF 根据第一请求中携带的参数确定的，所述第一请求结果包括接受第一请求或不接受第一请求；

其中，若第一请求中携带的参数中不包含模型不同层的时延需求信息，则模型不同层的时延需求信息由 AF 提供。

本公开实施例中，当 UE 和 AF 联合推理模型时，UE 基于自身能力进行模型切分点的判断，上报模型切分点（模型切分比例）给 AF，并进行联合推理交互。

可选地，所述确定 AI/ML 模型切分的切分结果之后，所述方法还包括：

通过 AF 向网络数据分析功能 NWDAF 发送第二请求，所述第二请求用于请求与所述 NWDAF 执行模型联合推理操作；其中，所述第二请求中携带的参数包括下述至少一项 AI/ML 模型切分点、与模型标识或模型切分标识关联的分析类型标识、接收 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足分析条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE 可提供的算力、待参与模型联合推理的 UE 可提供的内存、待参与模型联合推理的 UE 剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第二请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

接收 NWDAF 发送的第二请求结果，所述第二请求结果是由 NWDAF 根据第一请求中携带的参数确定的，所述第二请求结果包括接受第一请求或不接受第一请求；

其中，若第二请求中携带的参数中不包含模型不同层的时延需求信息，则模型不同层的时延需求信息由 AF 提供。

本公开实施例中，当 UE 和 NWDAF 联合推理模型时，UE 基于自身能力进行模型切分点的判断，上报 NWDAF，并进行联合推理交互。

第三方面，本公开提供一种辅助模型切分的装置，所述装置应用于网络实体，所述装置包括存储器，收发机，处理器：

- 5 存储器，用于存储计算机程序；收发机，用于在所述处理器的控制下收发数据；处理器，用于读取所述存储器中的计算机程序并执行以下操作：
接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息；
根据所述第一消息，确定 AI/ML 模型切分的切分结果。

- 10 第四方面，本公开提供一种辅助模型切分的装置，所述装置应用于用户设备 UE，所述装置包括存储器，收发机，处理器：

存储器，用于存储计算机程序；收发机，用于在所述处理器的控制下收发数据；处理器，用于读取所述存储器中的计算机程序并执行以下操作：
根据自身能力信息，确定 AI/ML 模型切分点信息；
将所述 AI/ML 模型切分点信息作为 AI/ML 模型切分的切分结果。

- 15 第五方面，本公开提供一种辅助模型切分的装置，所述装置应用于网络实体，所述装置包括：

接收单元，用于接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息；

确定单元，用于根据所述第一消息，确定 AI/ML 模型切分的切分结果。

- 20 第六方面，本公开提供一种辅助模型切分的装置，所述装置应用于用户设备 UE，所述装置包括：

确定单元，用于根据自身能力信息，确定 AI/ML 模型切分点信息；

处理单元，用于将所述 AI/ML 模型切分点信息作为 AI/ML 模型切分的切分结果。

- 25 第七方面，本公开提供一种处理器可读存储介质，所述处理器可读存储介质存储有计算机程序，所述计算机程序用于使所述处理器执行第一方面或第二方面任一项所述的方法。

- 30 本公开提供一种辅助模型切分的方法、装置及可读存储介质，通过接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息，并根据第一消息中的待参与模型联合推理的 UE 可提供的算力、待参与模型联合推理的 UE

可提供的内存、待参与模型联合推理的 UE 剩余的电量来确定切分 AI/ML 模型的模型切分点信息,如果推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,还可以结合推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息来确定切分 AI/ML 模型的模型切分点信息,作为切分结果,用以实现网络实体或待参与模型联合推理的 UE 基于切分结果对模型进行联合推理操作,因此,实现了基于终端(即 UE)能力对模型切分的分析,进而有效地实现对终端隐私的保护以及网络资源的优化。

应当理解,上述发明内容部分中所描述的内容并非旨在限定本公开的实施例的关键或重要特征,亦非用于限制本公开的范围。本公开的其它特征将通过以下的描述变得容易理解。

附图说明

为了更清楚地说明本公开或现有技术中的技术方案,下面将对实施例或现有技术描述中所需要使用的附图作一简单地介绍,显而易见地,下面描述中的附图是本公开的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动性的前提下,还可以根据这些附图获得其他的附图。

图 1 为本公开实施例提供的支持网络数据分析的 5GC 的网络架构图;

图 2 为本公开实施例提供的辅助模型切分的方法的第一流程示意图;

图 3 为本公开实施例一提供的当网络实体是 NWDAF 时辅助模型切分的方法的第一信令流程示意图;

图 4 为本公开实施例一提供的当网络实体是 NWDAF 时辅助模型切分的方法的第二信令流程示意图;

图 5 为本公开实施例一提供的当网络实体是 NWDAF 时辅助模型切分的方法的第三信令流程示意图;

图 6 为本公开实施例一提供的当网络实体是 NWDAF 时辅助模型切分的方法的第四信令流程示意图;

图 7 为本公开实施例二提供的当网络实体是 AF 时辅助模型切分的方法的第一信令流程示意图;

图 8 为本公开实施例二提供的当网络实体是 AF 时辅助模型切分的方法的第二信令流程示意图;

图 9 为本公开实施例三提供的当网络实体是 MMF 时辅助模型切分的方法的第一信令流程示意图；

图 10 为本公开实施例三提供的当网络实体是 MMF 时辅助模型切分的方法的第二信令流程示意图；

5 图 11 为本公开实施例四提供的当网络实体是 PCF 时辅助模型切分的方法的第一信令流程示意图；

图 12 为本公开实施例四提供的当网络实体是 PCF 时辅助模型切分的方法的第二信令流程示意图；

图 13 为本公开实施例提供的辅助模型切分的方法的第二流程示意图；

10 图 14 为本公开实施例提供的辅助模型切分的装置的结构示意图；

图 15 为本公开另一实施例提供的辅助模型切分的装置的结构示意图；

图 16 为本公开又一实施例提供的辅助模型切分的装置的结构示意图；

图 17 为本公开又一实施例提供的辅助模型切分的装置的结构示意图。

15 具体实施方式

本公开中术语“和/或”，描述关联对象的关联关系，表示可以存在三种关系，例如，A 和/或 B，可以表示：单独存在 A，同时存在 A 和 B，单独存在 B 这三种情况。字符“/”一般表示前后关联对象是一种“或”的关系。

下面将结合本公开实施例中的附图，对本公开实施例中的技术方案进行
20 清楚、完整地描述，显然，所描述的实施例仅仅是本公开一部分实施例，并不是全部的实施例。基于本公开中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本公开保护的范围。

为了清楚理解本公开的技术方案，首先对现有技术的方案进行详细介绍。现有技术中，在 SA#93e 通过的 SA1 R18 需求中，需要 AI/ML 模型传输的场
25 景如下：AI/ML 端点之间的 AI/ML 模型切分。

一个 AI/ML 模型可以基于当前的任务或者环境切分成多个部分。趋势是将计算复杂，能耗大的部分由网络进行推理，需要隐私保护或者时延敏感的部分在终端推理。例如，终端下载/机载一个模型，先推理特定几层/部分，然后把中间结果发送给网络；网络再执行剩余层/部分，然后把推理结果反馈给
30 终端。趋势是将计算复杂，能耗大的部分由网络进行推理，需要隐私保护或

者时延敏感的部分在终端推理。

该场景中，需要支持和辅助 AI/ML 模型用户（例如，application client running on the UE，即应用程序客户端运行在 UE 上）对模型的应用层本地训练，并支持将应用层推理反馈给 AI/ML 模型提供者（例如，应用功能（英文为：Application Function，简称为：AF））。

但是，现有技术中，无法实现基于终端能力对模型切分的分析，进而无法有效地实现对终端隐私的保护以及网络资源的优化。

发明人进一步研究发现，要基于终端能力对模型切分进行有效地分析，进而确定模型切分点，需要网络实体或者终端基于终端的电量、可提供的内存、可提供的算力等信息，以及关于此模型的时延需求以及关于此模型需要的推理数据集的隐私等级等，选择出模型切分点，然后网络实体或者终端将模型切分点信息发送给参与模型联合推理的网络实体或者终端实现基于终端能力对模型切分的分析，进而有效地实现对终端隐私的保护以及网络资源的优化。

所以基于上述发明人的创造性研究，提出了本公开提出的辅助模型切分的方法，本公开中，网络实体通过接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息，并根据第一消息中的待参与模型联合推理的 UE(s) 可提供的算力、待参与模型联合推理的 UE(s) 可提供的内存、待参与模型联合推理的 UE(s) 剩余的电量来确定切分 AI/ML 模型的模型切分点信息，如果推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，还可以结合推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息来确定切分 AI/ML 模型的模型切分点信息，作为切分结果，用以实现网络实体或待参与模型联合推理的 UE(s) 基于切分结果对模型进行联合推理操作，因此，实现了基于终端（即 UE）能力对模型切分的分析，进而有效地实现对终端隐私的保护以及网络资源的优化。结合图 1 所示，图 1 为本公开实施例提供的支持网络数据分析的 5GC 的网络架构图，如图 1 所示，本公开实施例中，网络数据分析功能（英文为：Network data analytic function，简称为：NWDAF）是运营商管理的网络分析功能，NWDAF 能够向 5G 核心网（英文为：5G Core Network，简称为：5GC）中各个网络功能（英文为：Network Function，简称为：NF）（即 NF(s)）、应用功能（英文为：Application Function，简称为：

AF) 和操作管理维护 (英文为: Operation Administration and Maintenance, 简称为: OAM) 提供数据分析服务。其中, 分析结果可以是历史统计信息或者预测信息。NWDAF 可以服务一个或多个网络切片。

其中, 在 5GC 中还包括其他多种功能。分别为用户平面功能(英文为: User Plane Function, 简称为: UPF)、会话管理功能(英文为: Session Management Function, 简称为: SMF)、接入和移动性管理功能 (英文为: Access and Mobility Management Function, 简称为: AMF)、统一数据库(英文为: Unified Data Repository, 简称为: UDR)、网络能力开放功能 (英文为: Network Exposure Function, 简称为: NEF)、AF、策略控制功能 (英文为: Policy Control Function, 简称为: PCF)及在线计费系统 (英文为: Online Charging System, 简称为: OCS)。其中, 这些其他功能均可统称为 NF。NWDAF 基于服务化接口与 5G 核心网中其他功能实体 5GC NF(s)及 OAM 进行通信。

5GC 中可以有不同 NWDAF 实例提供不同类型的专用分析。为了让消费者 NF 能够发现合适的 NWDAF 实例来提供特定类型的分析, NWDAF 实例需在向网络数据库功能(英文为: Network Repository Function, 简称为: NRF)注册时提供其支持 Analytic ID, Analytic ID 代表了分析类型标识 (这样消费者 NF 可以在向 NRF 查询 NWDAF 实例时, 提供 Analytic ID 来指示需要何种类型的分析。5GC 网络功能和 OAM 决定如何使用网络数据分析功能 NWDAF 提供的数据分析来提高网络性能。

本公开实施例中, 在一种应用场景中, 网络实体或者终端基于终端的电量、可提供的内存、可提供的算力等信息, 以及关于此模型的时延需求以及关于此模型需要的推理数据集的隐私等级等, 选择出模型切分点, 网络实体或者终端将模型切分点信息发送给参与模型联合推理的网络实体或者终端。

其中, 可提供的算力可以是终端的剩余运存。

因此, 网络实体或终端根据终端可提供的算力、终端可提供的内存、终端剩余的电量来确定切分 AI/ML 模型的模型切分点信息, 如果推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 还可以结合推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息来确定切分 AI/ML 模型的模型切分点信息, 作为切分结果, 用以实现网络实体或终端基于切分结果对模型进行联合推理操作, 因此, 实现了基于终端(即 UE(s))

能力对模型切分的分析，进而有效地实现对终端隐私的保护以及网络资源的优化。

需要说明的是，UE(s)可以为一个UE或多个UE(或一组UE或任意UEs)，其UE的数量可以根据具体场景确定。此外，下述实施例中针对UE上报的可提供的算力、可提供的内存、剩余的电量等场景，对于UE侧来说，是每个UE进行上报，其上报的参数为该UE自身的参数(比如该UE可提供的算力、可提供的内存、剩余的电量)；对于接收侧来说，接收每个UE上报的参数，即接收到的是各个UE上报的参数(结合来说，接收到的参数有各个UE(或UE(s))可提供的算力、可提供的内存、剩余的电量)。

以下将参照附图来描述本公开的实施例。

图2为本公开实施例提供的辅助模型切分的方法的第一流程示意图，如图2所示，本实施例提供的辅助模型切分的方法的执行主体为网络实体，该网络实体可以是NWDAF、AF、新的网络实体功能(例如，英文为：Model Management Function，简称为：MMF)功能、PCF中的任一个。本公开实施例提供的辅助模型切分的方法包括以下步骤：

步骤101、网络实体接收用于辅助人工智能/机器学习AI/ML模型切分分析的第一消息。

可选地，所述第一消息中包括下述至少一项：待参与AI/ML模型切分的UE(s)或者用户永久标识SUPI、使用AI/ML模型的应用的标识、待参与模型联合推理的UE(s)可提供的算力、待参与模型联合推理的UE(s)可提供的内存、待参与模型联合推理的UE(s)剩余的电量、AI/ML模型的大小；

若推理AI/ML模型时需要数据集的隐私等级和/或模型不同层的时延，所述第一消息中还包括：推理AI/ML模型需要的数据集的隐私等级和/或模型不同层的时延需求信息。

其中，接收第一消息的具体方式可以依据具体的网络实体在相应的场景下确定，在此不做具体地限定。

步骤102、网络实体根据所述第一消息，确定AI/ML模型切分的切分结果。

可选地，所述根据所述第一消息，确定AI/ML模型切分的切分结果，可以通过以下步骤实现：

步骤 a1、根据所述第一消息中的待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量，确定所述待参与模型联合推理的 UE(s)对应的模型切分点信息；

5 步骤 a2、若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，则根据所述第一消息中的待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量、推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息，确定所述待参与模型联合推理的 UE(s)对应的模型切分点信息；

10 步骤 a3、将所述模型切分点信息作为所述切分结果。

具体地，网络实体可以基于终端的电量、可提供的内存、可提供的算力等信息，以及关于此模型的时延需求以及关于此模型需要的推理数据集的隐私等级等，选择出模型切分点（即模型切分点信息或 AI/ML 模型切分点信息），然后网络实体将模型切分点信息发送给参与模型联合推理的网络实体或者终端。

15 可选地，所述确定所述待参与模型联合推理的 UE(s)对应的模型切分点信息，可以通过以下步骤实现：

步骤 b1、将所有确定参与执行模型联合推理的 UE 中的任一 UE 作为目标 UE，针对每个所述目标 UE 执行下述步骤：

20 步骤 b2、若所述目标 UE 可提供的算力低于第一预设算力阈值或所述目标 UE 可提供的内存低于第一预设内存阈值或所述目标 UE 剩余的电量低于第一预设电量阈值时，确定所述目标 UE 执行第一预设数目层的推理；

步骤 b3、若所述目标 UE 可提供的算力高于第一预设算力阈值并低于第二预设算力阈值，和所述目标 UE 可提供的内存高于第一预设内存阈值并低于第二预设内存阈值，以及所述目标 UE 剩余的电量高于第一预设电量阈值并低于第二预设电量阈值时，确定所述目标 UE 执行第二预设数目的推理，第一预设数目层小于第二预设数目层；

25 步骤 b4、若所述目标 UE 可提供的算力高于第二预设算力阈值并低于第三预设算力阈值，和所述目标 UE 可提供的内存高于第二预设内存阈值并低于第三预设内存阈值，以及所述目标 UE 剩余的电量高于第二预设电量阈值

并低于第三预设电量阈值时,确定所述目标 UE 执行第三预设数目层的推理,第二预设数目层小于第三预设数目层;以此类推,直至确定所述目标 UE 执行第 N+1 预设数目层的推理,第 N 预设数目层小于第 N+1 预设数目层;

5 步骤 b5、若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,且若推理 AI/ML 模型需要的数据集的隐私等级低于第一预设隐私等级或模型不同层的时延需求信息低于第一预设时延,则所述目标 UE 执行所述第一预设数目层的推理;

10 步骤 b6、若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,且若推理 AI/ML 模型需要的数据集的隐私等级高于第一预设隐私等级并低于第二预设隐私等级,和模型不同层的时延需求信息高于第一预设时延并低于第二预设时延时,确定所述目标 UE 执行第二预设数目层的推理;

15 步骤 b7、若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,且若推理 AI/ML 模型需要的数据集的隐私等级高于第二预设隐私等级并低于第三预设隐私等级,和模型不同层的时延需求信息高于第二预设时延并低于第三预设时延时,确定所述目标 UE 执行第三预设数目层的推理;以此类推,直至确定所述目标 UE 执行第 N+1 预设数目层的推理,第 N 预设数目层小于第 N+1 预设数目层;

20 步骤 b8、根据预设数目层确定 AI/ML 模型切分点信息,所述 AI/ML 模型切分点信息用于表示 AI/ML 模型切分比例,所述预设数目层包括第 N 预设数目层,N 大于或等于一。

具体地,参与模型联合推理的每个 UE 都需要确定其推理的层数或部分,针对每个 UE (即目标 UE) 可以通过以下过程实现切分点的确定:

25 目标 UE 的可提供的算力和内存较少、剩余的电量较低时,只能做较少层的推理,如果推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,且隐私等级需求较高和/或模型不同层的时延需求较低时做较少层的推理;如果 UE 的可提供的算力和内存充足、剩余的电量较高时,可以做较多层的推理,如果推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,且隐私等级需求较低和/或模型不同层的时延需求较高时做较多层的推理。例如,假设对于一个 8 层的 AI/ML 模型,UE1 可以做两层的推理(比如,30 UE1 做前两层的推理)、UE2 可以做三层的推理(比如,UE2 做中间三层的

推理), AF 做最后三层的推理, 即 UE1 对应模型切分点 1, UE2 对应模型切分点 2。

本实施例中, 网络实体通过接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息, 并根据第一消息中的待参与模型联合推理的 UE(s) 可提供的算力、待参与模型联合推理的 UE(s) 可提供的内存、待参与模型联合推理的 UE(s) 剩余的电量来确定切分 AI/ML 模型的模型切分点信息, 如果推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 还可以结合推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息来确定切分 AI/ML 模型的模型切分点信息, 作为切分结果, 用以实现网络(AF) 和待参与模型联合推理的 UE(s) 基于切分结果对模型进行联合推理, 因此, 实现了基于终端(即 UE) 能力对模型切分的分析, 进而有效地实现对终端隐私的保护以及网络资源的优化。

示例性地, 实施例一(网络实体为 NWDAF, NWDAF 进行模型切分点的判断), 下述以至少四种场景为例对辅助模型切分的方法进行详细说明。

场景 11: 当 UE(s) 和 AF(s) 联合推理模型(s) 时, AF 请求 NWDAF 收集 UE 的能力进行模型切分点的判断, NWDAF 收集并判断后将结果分别反馈给 AF, AF 发给相关的 UE。

可选地, 如果所述网络实体是网络数据分析功能 NWDAF, 所述接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息, 可以包括以下步骤:

步骤 c11、直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第一请求, 所述第一请求用于请求对 AI/ML 模型切分进行分析; 其中, 所述第一请求中携带的参数包括下述至少一项: 与模型标识或模型切分标识关联的分析类型标识、接受 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足切分条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE(s) 可提供的算力、待参与模型联合推理的 UE(s) 可提供的内存、待参与模型联合推理的 UE(s) 剩余的电量; 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 所述第一请求中携带的参数还包括: 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延

需求信息;

步骤c12、根据所述第一请求中携带的参数,向5GC NF(s)发送第二消息,所述第二消息用于请求5GC NF(s)采集UE(s)对应的第一数据,所述第一数据中包括下述至少一项:待参与AI/ML模型切分的UE(s)以及SUPI、待参与模型联合推理的UE(s)可提供的算力、待参与模型联合推理的UE(s)可提供的内存、待参与模型联合推理的UE(s)剩余的电量、AI/ML模型的大小;若第一请求中携带的参数包括推理AI/ML模型需要的数据集的隐私等级和/或模型不同层的时延需求信息,则所述第一数据中还包括:推理AI/ML模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

步骤c13、接收5GC NF(s)发送的所述第一数据,并将所述第一数据作为所述第一消息;所述第一数据是由待参与模型联合推理的UE(s)同意所述第二消息的请求后提供的。

可选地,所述确定AI/ML模型切分的切分结果之后,所述方法还包括:向所述AF发送第三消息,所述第三消息中包括AI/ML模型切分点信息以及所述第一数据;所述第三消息用于为AF提供向待参与模型联合推理的UE(s)发送第二请求时携带的参数,所述第二请求用于请求与待参与模型联合推理的UE(s)执行模型联合推理操作,所述第二请求中携带的参数包括所述第三消息。

具体地,参见图3,图3为本公开实施例一提供的当网络实体是NWDAF时辅助模型切分的方法的第一信令流程示意图。具体步骤为:(301a是AF在受信任区域,301b、301c是AF在不受信任区;步骤305a是AF在受信任区域,305b,305c是AF在不受信任区。)

步骤301a,AF向NWDAF发送Nnwdaf_MLModelSplit_Request(即机器学习模型切分请求),请求中包含和Model ID关联的Analytics ID(即分析类型标识)=MLModelSplit,请求NWDAF收集UE(即待参与模型联合推理的UE(s))的可提供的算力、可提供的内存、剩余的电量,以及推理此模型需要的数据集的隐私等级,以及可选的包含模型不同层的时延需求等信息,如下表1:

表1

Parameter Name (即参数名)	Parameter Value (即参数值)	Parameter Description (即参数描述)
Analytics ID	定义新的 Analytics ID, 例如 MLModelSplit	AI/ML 模型切分
Target of Analytics Reporting	UE ID, UE group ID, or “any UE”	接受 AI/ML 模型切分的一个 UE 或一组 UEs 的标识, 或满足切分条件(例如 AoI 内)的任意 UEs
AoI (Area of Interest)		进行 AI/ML 模型切分的区域
Model size		AI/ML 模型的大小
UE 可提供的算力		终端可提供的算力
UE 可提供的内存		终端可提供的内存
UE 剩余的电量		终端剩余的电量
如果有, 数据集的隐私等级		推理此模型需要的数据集的隐私等级(如果 SA3 可以提供隐私等级设置)
模型不同层的时延需求		模型不同层的时延需求信息

步骤 301b, AF 向 NEF 发送 Nnef_MLModelSplit_Request (即机器学习模型切分请求), 请求中包含上述表 1 中的信息。

步骤 301c, NEF 授权后, 向 NWDAF 发送 Nnwdafe_MLModelSplit_Request, 请求中包含上述表 1 中的信息。

5 步骤 302, NWDAF 调用 Nnf_EventExposure_Subscribe (即事件开放订阅)

- 向 5GC NF(s) (例如, AMF/SMF) 采集 UE(s)的可提供的算力、可提供的内存、剩余的电量, (如果有)推理此模型需要的数据集的隐私等级, 以及(如果能提供)模型不同层的时延需求; 5GC NF(s) (例如, AMF/SMF) 将这些请求信息发送给终端, 终端收到这个请求, (如果终端通过这个请求)就准备和
- 5 该 Analytics ID (Model ID) 对应的电量、内存、算力, (如果有)数据集的隐私等级等信息反馈 (参见表 2) 给 NF(例如, AMF/SMF)。

表 2

Information (即数据)	Source (即来源)	Description (即数据描述)
UE ID	AMF	可参与 AI/ML 模型切分的 UE(s), SUPI
Application ID	SMF, AF	使用 AI/ML 模型的应用的标识
UE 可提供的算力	AMF, SMF	终端可提供的算力
UE 可提供的内存	AMF, SMF	终端可提供的内存
UE 剩余的电量	AMF, SMF	终端剩余的电量
如果有, 数据集的隐私等级	AMF, SMF	推理此模型需要的数据集的隐私等级
模型不同层的时延需求	AF, SMF	模型不同层的时延需求信息
Model(s) size	AF, SMF	AI/ML 模型的大小

步骤 303, 5GC NF(s)调用 Nnf_EventExposure_Notify (即事件开放通知) 向 NWDAF 反馈所需数据。

- 10 步骤 304, NWDAF 执行分析, 基于 UE 的可提供的算力、可提供的内存、剩余的电量, 以及推理此模型需要的数据集的隐私等级, 以及模型不同层的时延需求, 选出模型切分点 (或者模型切分比例)。

例如, UE 的可提供的算力和内存较少、剩余的电量较低时, 只能做较少

层的推理,如果推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,且隐私等级需求较高和/或模型不同层的时延需求较低时做较少层的推理;如果 UE 的可提供的算力和内存充足、剩余的电量较高时,可以做较多层的推理,如果推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,且隐私等级需求较低和/或模型不同层的时延需求较高时做较多层的推理;例如,假设对于一个 8 层的 ML 模型,UE1 可以做前两层的推理、UE2 做中间三层的推理,AF 做最后三层的推理,即 UE1 对应模型切分点 1,UE2 对应模型切分点 2。

305a, 通过 Nnwdaf_MLModelSplit_Request Response (即机器学习模型切分请求的响应) 将模型切分点 (或者模型切分比例) 等信息 (如下表 3) 发送给 AF。

表 3

Information	Description
List of UE(s)	可参与 AI/ML 模型切分的 UE(s), SUPI
List of Application(s)	使用 AI/ML 模型的应用(s)的标识
UE 可提供的算力	终端可提供的算力
UE 可提供的内存	终端可提供的内存
UE 剩余的电量	终端剩余的电量
如果有, 数据集的隐私等级	推理此模型需要的数据集的隐私等级
模型不同层的时延需求	模型不同层的时延需求信息
Model split point(s)	(一个或多个)模型切分点 (或者模型切分比例)

步骤 305b, NWDAF 向 NEF 发送 Nnwdaf_MLModelSplit_Request Response, 内容包含上述模型切分点 (或者模型切分比例) 等的信息。

步骤 305c, NEF 授权后, 向 NWDAF 发送 Nnef_MLModelSplit_Request response (即机器学习模型切分请求的响应), 内容包含上述模型切分点 (或

者模型切分比例)等的信息。

步骤 306, AF 和相关 UE(s) 建立连接。

步骤 307, AF 向相关 UE(s) 发送模型联合推理请求 Naf_MLModelJointInference_Request (即模型联合推理请求), 请求中包含模型切分点 (或者模型切分比例) 信息。

步骤 308, UE(s) 向 AF 发送模型联合推理请求的响应 Naf_MLModelJointInference_Request response (即模型联合推理请求的响应), 表示是否接受这个模型联合推理请求。

场景 12: 当 UE(s) 和 AF(s) 联合推理 (多个) 模型时, UE 上报自身的能力给 AF, AF 请求 NWDAF 进行模型切分点的判断, NWDAF 把判断结果反馈给 AF, AF 发送给 UE。

可选的, 如果所述网络实体是网络数据分析功能 NWDAF, 所述接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息, 可以通过包括以下步骤:

直接或通过网络能力开放功能 NEF 接收 AF 发送的第三请求; 所述第三请求是由 AF 确定与待参与模型联合推理的 UE(s) 具备完全支持执行模型联合推理的能力后确定的, 其中, 是否具备完全支持执行模型联合推理的能力是由 AF 根据接收到的待参与模型联合推理的 UE(s) 发送的第四请求中携带的参数确定的, 所述第四请求用于待参与模型联合推理的 UE(s) 请求与 AF 进行模型联合推理, 所述第四请求中携带的参数包括下述至少一项: 与模型标识或模型切分标识关联的分析类型标识、待参与模型联合推理的 UE(s) 可提供的算力、待参与模型联合推理的 UE(s) 可提供的内存、待参与模型联合推理的 UE(s) 剩余的电量; 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 所述第四请求中携带的参数还包括: 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息, 所述第三请求中携带的参数包括所述第四请求中携带的参数; 将所述第三请求中携带的参数作为所述第一消息;

其中, 若第四请求中携带的参数中不包含模型不同层的时延需求信息, 则模型不同层的时延需求信息由 AF 提供。

可选地，所述确定 AI/ML 模型切分的切分结果之后，所述方法还包括：
向所述 AF 发送第四消息，所述第四消息中包括所述 AI/ML 模型切分点、
所述第一消息中的数据；所述第四消息用于为 AF 提供向待参与模型联合推理的 UE(s) 发送的第四请求结果，所述第四请求结果中包括所述 AI/ML 模型
5 切分点信息。

具体地，参见图 4，图 4 为本公开实施例一提供的当网络实体是 NWDAF 时辅助模型切分的方法的第二信令流程示意图。具体步骤为：（404a 是 AF 在受信任区域，404b、404c 是 AF 在不受信任区；步骤 405a 是 AF 在受信任区域，405b、405c 是 AF 在不受信任区。）

10 步骤 400，AF 和 UE 建立连接

步骤 401，UE 发起和 AF 进行模型联合推理请求 Naf_MLModelJointInference_Request，请求中包含和 Model ID 关联的 Analytics ID=MLModelSplit、UE 自身的可提供的算力、可提供的内存、剩余的电量等信息；如果有推理 AI/ML 模型时需要数据集的隐私等级，则请求中还包含推理
15 推理 AI/ML 模型需要的数据集的隐私等级，如果没有，基于 SA3 确认是否可以设定；如果有推理 AI/ML 模型时需要模型不同层的时延需求信息，则请求中还包含模型不同层的时延需求信息，如果没有，由 AF 提供。

步骤 402，AF 判断这个联合推理请求。

步骤 403a，AF 向 NWDAF 发送 Nnwdaf_MLModelSplit_Request，请求
20 中包含和 Model ID 关联的 Analytics ID=MLModelSplit，UE 的可提供的算力、可提供的内存、剩余的电量，以及推理此模型需要的数据集的隐私等级，以及模型不同层的时延需求等信息，并请求 NWDAF 进行模型切分点的判断。

步骤 403b，AF 向 NEF 发送 Nnef_MLModelSplit_Request，请求中包含上述信息。

25 步骤 403c，NEF 授权后，向 NWDAF 发送 Nnwdaf_MLModelSplit_Request，请求中包含上述信息。

步骤 404，NWDAF 执行分析，基于 UE 的可提供的算力、可提供的内存、剩余的电量，以及推理此模型需要的数据集的隐私等级，以及模型不同层的时延需求，选出模型切分点（或者模型切分比例）。

30 例如，UE 的可提供的算力、可提供的内存、剩余的电量较低时，只能做

较少层的推理,如果推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,且隐私等级需求较高和/或模型不同层的时延需求较低时做较少层的推理;如果 UE 的可提供的算力和内存充足、剩余的电量较高时,可以做较多层的推理,如果推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,且隐私等级需求较低和/或模型不同层的时延需求较高时做较多层的推理。例如,假设对于一个 8 层的 ML 模型,UE1 可以做前两层的推理、UE2 做中间三层的推理,AF 做最后三层的推理,即 UE1 对应模型切分点 1,UE2 对应模型切分点 2。

步骤 405a, NWDAF 通过 Nnwdaf_MLModelSplit_Request Response 将模型切分点 (或者模型切分比例) 等信息发送给 AF。

步骤 405b, NWDAF 向 NEF 发送 Nnwdaf_MLModelSplit_Request Response, 内容包含上述模型切分点 (或者模型切分比例)。

步骤 505c, NEF 授权后, 向 NWDAF 发送 Nnef_MLModelSplit_Request response, 内容包含上述模型切分点 (或者模型切分比例)。

步骤 406, AF 把模型联合推理请求的响应 Naf_MLModelJointInference_Request response 发送给 UE, 响应中包含模型切分点 (或者模型切分比例)。

场景 13、当 UE(s)和 AF(s)联合推理模型时, UE 上报自身的能力给 AF, AF 请求 NWDAF 进行模型切分点的判断并请求其他可参与模型推理的 UE(s)参与本模型的推理, NWDAF 把收集和判断结果反馈给 AF, AF 发送给 UE。

可选地, 如果所述网络实体是网络数据分析功能 NWDAF, 所述接收用于请求人工智能/机器学习 AI/ML 模型切分分析的第一消息, 可以包括以下步骤:

步骤 c31、直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第五请求, 所述第五请求是由 AF 确定与待参与模型联合推理的 UE(s)不具备完全支持执行模型联合推理的能力后确定的, 其中, 是否具备完全支持执行模型联合推理的能力是由 AF 根据接收到的待参与模型联合推理的 UE(s)发送的第六请求确定的, 所述第五请求用于请求对 AI/ML 模型切分进行分析以及请求查找其他可参与模型推理的 UE(s)并提供其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量; 其中, 所述第五请求中携带的参

数包括下述至少一项：与模型切分关联的分析类型标识、接受 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足切分条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第五请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；所述第六请求用于待参与模型联合推理的 UE(s)请求与 AF 进行模型联合推理，所述第六请求中携带的参数包括所述第五请求中携带的参数；

步骤 c32、根据所述第五请求中携带的参数，向 5GC NF(s)发送第五消息，所述第五消息用于请求 5GC NF(s)采集 UE(s)对应的第二数据，所述第二数据中包括下述至少一项：待参与 AI/ML 模型切分的 UE(s)或者 SUPI、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量、AI/ML 模型的大小；若第五请求中携带的参数包括推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息，则所述第二数据中还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

步骤 c33、接收 5GC NF(s)发送的所述第二数据，并将所述第二数据作为所述第一消息；所述第二数据是由待参与模型联合推理的 UE(s)同意所述第五消息的请求后提供的；

其中，若第六请求中携带的参数中不包含模型不同层的时延需求信息，则模型不同层的时延需求信息由 AF 提供。

可选地，所述确定 AI/ML 模型切分的切分结果之后，所述方法还包括：向所述 AF 发送第六消息，所述第六消息中至少包括所述 AI/ML 模型切分点信息以及所述第二数据；所述第六消息用于为 AF 提供向待参与模型联合推理的 UE(s)发送的第六请求结果，所述第六请求结果中包括所述待参与模型联合推理的 UE(s)对应的所述 AI/ML 模型切分点信息、所述其他可参与模型推理的 UE(s)以及对应的所述 AI/ML 模型切分点信息。

具体地，参见图 5，图 5 为本公开实施例一提供的当网络实体是 NWDAF 时辅助模型切分的方法的第三信令流程示意图。具体步骤为：（步骤 503a 是

AF 在受信任区域, 步骤 503b、步骤 503c 是 AF 在不受信任区; 步骤 507a 是 AF 在受信任区域, 步骤 507b、步骤 507c 是 AF 在不受信任区。)

步骤 500-501 同场景 12 的实施例中的步骤 400-401 所描述。(即:
步骤 500, AF 和 UE 建立连接。

5 步骤 501, UE 发起和 AF 进行模型联合推理请求 Naf_MLModelJointInference_Request, 请求中包含和 Model ID 关联的 Analytics ID=MLModelSplit, UE 自身的可提供的算力、可提供的内存、剩余的电量等信息; 如果有推理 AI/ML 模型时需要数据集的隐私等级, 则请求中还包含推理 AI/ML 模型需要的数据集的隐私等级, 如果没有, 基于 SA3 确认是否可以
10 设定; 如果有推理 AI/ML 模型时需要模型不同层的时延需求信息, 则请求中还包含模型不同层的时延需求信息, 如果没有, 由 AF 提供。步骤 502, AF 判断联合推理请求, 如果判断 UE 和 AF 自身能力不够做联合推理, 则认为需要其他 UE(s) 帮忙进行模型切分。

 步骤 503a, AF 向 NWDAF 发送 Nnwdaf_MLModelSplit_Request, 请求
15 中包含和 Model ID 关联的 Analytics ID=MLModelSplit, UE 的可提供的算力、可提供的内存、剩余的电量, 以及推理此模型需要的数据集的隐私等级, 以及模型不同层的时延需求等信息, 以及请求 NWDAF 帮忙发现其他可参与模型推理的 UE(s) 以及其相关的可提供的算力、可提供的内存、剩余的电量等并请求 NWDAF 进行模型切分点的判断。

20 步骤 503b, AF 向 NEF 发送 Nnef_MLModelSplit_Request, 请求中包含上述信息。

 步骤 503c, NEF 授权后, 向 NWDAF 发送 Nnwdaf_MLModelSplit_Request, 请求中包含上述信息。

 步骤 504-步骤 505 同场景 11 的实施例中的步骤 302-303 所描述(即:
25 步骤 504, NWDAF 调用 Nnf_EventExposure_Subscribe 向 5GC NF(s) (例如, AMF/SMF) 采集 UE(s) 的可提供的算力、可提供的内存、剩余的电量, (如果有) 推理此模型需要的数据集的隐私等级, 以及(如果能提供) 模型不同层的时延需求; 5GC NF(s) (例如, AMF/SMF) 将这些请求信息发送给终端, 终端收到这个请求, (如果通过这个请求) 就准备和该 Analytics ID (Model
30 ID) 对应的电量、内存、算力, (如果有) 数据集的隐私等级等信息反馈给

NF(例如, AMF/SMF)。

步骤 505, 5GC NF(s)调用 Nnf_EventExposure_Notify 向 NWDAF 反馈所需数据)。

5 步骤 506, NWDAF 执行分析, 并基于 output data 中的可参与模型切分的列表中的 UE(s), 以及和他们对应的可提供的算力、可提供的内存、剩余的电
量, 以及推理此模型需要的数据集的隐私等级, 以及模型不同层的时延需求,
选出模型切分点 (或者模型切分比例)。

例如, UE 的可提供的算力和内存较少、剩余的电量较低时, 只能做较少
层的推理, 如果推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的
10 时延, 且隐私等级需求较高和/或模型不同层的时延需求较低时做较少层的推
理; 如果 UE 的可提供的算力和内存充足、剩余的电量较高时, 可以做较多
层的推理, 如果推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的
时延, 且隐私等级需求较低和/或模型不同层的时延需求较高时做较多层的推
理; 例如, 假设对于一个 8 层的 ML 模型, UE1 可以做前两层的推理、UE2
15 做中间三层的推理, AF 做最后三层的推理, 即 UE1 对应模型切分点 1, UE2
对应模型切分点 2。

步骤 507 (即步骤 507a-507c) 同场景 11 的实施例中的步骤 305 (即步骤
305a-305c) 所描述 (即:

20 步骤 507a, 通过 Nnwdaf_MLModelSplit_Request Response 将模型切分点
(或者模型切分比例) 等信息发送给 AF, 如表 3。

步骤 507b, NWDAF 向 NEF 发送 Nnwdaf_MLModelSplit_Request Response,
内容包含上述模型切分点 (或者模型切分比例) 等的信息。

步骤 507c, NEF 授权后, 向 NWDAF 发送 Nnef_MLModelSplit_Request
response, 内容包含上述模型切分点 (或者模型切分比例) 等的信息)。

25 步骤 508, AF 把模型联合推理请求的响应
Naf_MLModelJointInference_Request response 发送给 UE, 响应中包含模型切
分点 (或者模型切分比例) 以及其他可参与模型推理的 UE(s) 以及其相关的模
型切分点 (或者模型切分比例) 信息。

30 场景 14: 当 UE 和 NWDAF 联合推理模型时, UE 通过 AF 上报自身的能力
NWDAF, 同时请求 NWDAF 进行模型切分点的判断, NWDAF 把判断结

果通过 AF 反馈给 UE。

可选地，如果所述网络实体是网络数据分析功能 NWDAF，所述接收用于请求人工智能/机器学习 AI/ML 模型切分分析的第一消息，可以包括以下步骤：

- 5 步骤 c41、接收待参与模型联合推理的 UE(s)发送的第七请求，所述第七请求用于请求与 NWDAF 执行模型联合推理操作；所述第七请求中携带的参数包括下述至少一项：待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第七请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；根据所述第七请求中携带的参数，确定 AF 与所述待参与模型联合推理的 UE(s)是否具备完全支持执行模型联合推理的能力；

- 15 步骤 c42、若与所述待参与模型联合推理的 UE(s)具备完全支持执行模型联合推理的能力，则将所述第七请求中携带的参数作为所述第一消息待参与 AI/ML 模型切分的 UE(s)待参与模型联合推理的 UE(s)待参与模型联合推理的 UE(s)待参与模型联合推理的 UE(s)；

- 20 步骤 c43、若与所述待参与模型联合推理的 UE(s)不具备完全支持执行模型联合推理的能力，则请求 NF 查找其他可参与模型推理的 UE(s)并提供其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量，且若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，则请求提供其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

- 25 步骤 c44、接收 NF 发送的查找结果，所述查找结果包括其他可参与模型推理的 UE(s)以及其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量，若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，则查找结果还包括：其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

- 30 步骤 c45、将所述第七请求中携带的参数以及所述查找结果作为所述第一消息；

其中，若第七请求中携带的参数中不包含模型不同层的时延需求信息，则模型不同层的时延需求信息由 AF 提供。

可选地，所述确定 AI/ML 模型切分的切分结果之后，所述方法还包括：

向所述待参与模型联合推理的 UE(s)发送第七请求结果，所述第七请求结果中包括 AI/ML 模型切分点信息；

其中，若存在其他参与联合模型推理的 UE(s)，则所述第七请求结果中还包括其他参与联合模型推理的 UE(s)对应的模型切分点信息，并通过 AF 透传给所述 UE。

具体地，参见图 6，图 6 为本公开实施例一提供的当网络实体是 NWDAF 时辅助模型切分的方法的第四信令流程示意图。具体步骤为：

步骤 600，AF 和 UE 建立连接。

步骤 601，UE 发起和 NWDAF 进行模型联合推理请求 NnwdaflMLModelJointInference_Request（即模型联合推理请求），请求中包含和 Model ID 关联的 Analytics ID= MLModelSplit，UE 自身的可提供的算力、可提供的内存、剩余的电量等信息；如果有推理 AI/ML 模型时需要数据集的隐私等级，则请求中还包含推理 AI/ML 模型需要的数据集的隐私等级，如果没有，基于 SA3 确认是否可以设定；如果有推理 AI/ML 模型时需要模型不同层的时延需求信息，则请求中还包含模型不同层的时延需求信息，如果没有，由 AF 提供；通过 AF 透传给 NWDAF。

步骤 602，NWDAF 判断联合推理请求。

可选地，步骤 603-604，如果需要，通过 NF 发现其他可参与模型推理的 UE(s)以及其相关的可提供的算力、可提供的内存、剩余的电量等。（即：

步骤 603，NWDAF 调用 Nnf_EventExposure_Subscribe 向 5GC NF(s)（例如，AMF/SMF）采集 UE(s)的可提供的算力、可提供的内存、剩余的电量，(如果有)推理此模型需要的数据集的隐私等级，以及(如果能提供)模型不同层的时延需求；5GC NF(s)（例如，AMF/SMF）将这些请求信息发送给终端，终端收到这个请求，（如果通过这个请求）就准备和该 Analytics ID（Model ID）对应的电量、内存、算力，（如果有）数据集的隐私等级等信息反馈给 NF(例如，AMF/SMF)。

步骤 604，5GC NF(s)调用 Nnf_EventExposure_Notify 向 NWDAF 反馈所

需数据。)

步骤 605, NWDAF 基于收集到的信息以及自身拥有的模型执行分析, 进行模型切分点 (或者模型切分比例) 的判断。

步骤 606, NWDAF 向 UE 反馈 AI/ML 模型联合推理请求的响应
5 NnwdaflMLModelJointInference_Request response (即模型联合推理请求的响应), 请求中包含模型切分点 (或者模型切分比例) 信息, 如果有其他参与联合模型推理的 UE(s), 则包含相应 UE(s) 的模型切分点 (或者模型切分比例) 信息, 通过 AF 透传给 UE。

示例性地, 实施例二 (网络实体是 AF, AF 进行模型切分点的判断。),
10 下述以至少两种场景为例对辅助模型切分的方法进行详细说明。

场景 21: 当 UE 和 AF 联合推理模型时, UE 上报自身的能力给 AF, AF 进行联合推理和模型切分点的判断, 然后把判断结果反馈给 UE。

可选地, 如果所述网络实体是应用功能 AF, 所述接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息, 可以包括以下步骤:

15 步骤 d11、接收待参与模型联合推理的 UE(s) 发送的第八请求, 所述第八请求用于请求与 AF 执行模型联合推理操作; 所述第八请求中携带的参数包括下述至少一项: 与模型标识或模型切分标识关联的分析类型标识、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE(s) 可提供的算力、待参与模型联合推理的 UE(s) 可提供的内存、待参与模型联合推理的
20 UE(s) 剩余的电量; 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 所述第八请求中携带的参数还包括: 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息; 根据所述第八请求中携带的参数, 确定与所述待参与模型联合推理的 UE(s) 是否具备完全支持执行模型联合推理的能力;

25 步骤 d12、若与所述待参与模型联合推理的 UE(s) 具备完全支持执行模型联合推理的能力, 则将所述第八请求中携带的参数作为所述第一消息;

步骤 d13、若与所述待参与模型联合推理的 UE(s) 不具备完全支持执行模型联合推理的能力, 则向网络数据分析功能 NWDAF 发送第九请求, 所述第九请求中携带的参数包括所述第八请求中携带的参数, 且所述第九请求用于

请求 NWDAF 通过 NF 查找其他可参与模型推理的 UE(s) 以及其他可参与模型推理的 UE(s) 可提供的算力、可提供的内存、剩余的电量, 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 所述第九请求中携带的参数还包括: 其他可参与模型推理的 UE(s) 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息; (其中, 所述第九请求结果是由 NWDAF 通过向 5GC NF(s) 采集 UE(s) 对应的第三数据确定的; 所述第三数据中包括下述至少一项: 参与 AI/ML 模型切分的 UE 或者 SUPI、待参与模型联合推理的 UE(s) (这里的待参与模型联合推理的 UE(s) 为上述的其他可参与模型推理的 UE(s)) 可提供的算力、待参与模型联合推理的 UE(s) (这里的待参与模型联合推理的 UE(s) 为上述的其他可参与模型推理的 UE(s)) 可提供的内存、待参与模型联合推理的 UE(s) (这里的待参与模型联合推理的 UE(s) 为上述的其他可参与模型推理的 UE(s)) 剩余的电量、AI/ML 模型的大小; 若第九请求中携带的参数包括推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息, 则所述第三数据中还包括: 其他可参与模型推理的 UE(s) 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息; 所述第九请求结果包括下述至少一项: 第三数据、其他可参与模型推理的 UE(s) 以及其他可参与模型推理的 UE(s) 可提供的算力、可提供的内存、剩余的电量, 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 所述第九请求结果还包括: 其他可参与模型推理的 UE(s) 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息。)

步骤 d14、接收 NWDAF 发送的第九请求结果, 并将所述第九请求结果以及所述第八请求中携带的参数作为所述第一消息; 其中, 所述第九请求结果包括其他可参与模型推理的 UE(s) 以及其他可参与模型推理的 UE(s) 可提供的算力、可提供的内存、剩余的电量, 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 所述第九请求结果还包括: 其他可参与模型推理的 UE(s) 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

若第八请求中携带的参数中不包含模型不同层的时延需求信息, 则模型不同层的时延需求信息由 AF 提供。

可选地, 在确定 AI/ML 模型切分的切分结果之后, 所述方法还包括:

将确定与所述待参与模型联合推理的 UE(s)是否具备完全支持执行模型联合推理的能力的结果、所述待参与模型联合推理的 UE(s)的相关信息以及所述待参与模型联合推理的 UE(s)对应的模型切分点信息发送给所述 UE;

其中, 所述相关信息包括下述至少一项: 待参与模型联合推理的 UE(s) 5 可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量; 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 所述相关信息中还包括: 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息。

具体地, 参见图 7, 图 7 为本公开实施例二提供的当网络实体是 AF 时辅助模型切分的方法的第一信令流程示意图。具体步骤为: (步骤 703a 的是 AF 在受信任区域, 703b、703c 是 AF 在不受信任区; 步骤 706a 是 AF 在受信任区域, 706b、706c 是 AF 在不受信任区。)

步骤 700-701 同场景 13 的实施例中的步骤 500-501 所描述 (即:

步骤 700, AF 和 UE 建立连接。

15 步骤 701, UE 发起和 AF 进行模型联合推理请求 Naf_MLModelJointInference_Request, 请求中包含和 Model ID 关联的 Analytics ID=MLModelSplit, UE 自身的可提供的算力、可提供的内存、剩余的电量等信息; 如果有推理 AI/ML 模型时需要数据集的隐私等级, 则请求中还包含推理 AI/ML 模型需要的数据集的隐私等级, 如果没有, 基于 SA3 确认是否可以 20 设定; (如果有推理 AI/ML 模型时需要模型不同层的时延需求信息, 则请求中还包含模型不同层的时延需求信息, 如果没有, 由 AF 提供。

步骤 702, AF 判断联合推理请求, 如果需要其他 UE(s)参与联合推理模型, 则执行步骤 703 (即步骤 703a-703c) -706, 否则跳过。

25 步骤 703 (即步骤 703a-703c) -706 (即步骤 706a-706c) 如场景 13 的实施例中的步骤 503 (即步骤 503a-503c) -505、步骤 507 (即步骤 507a-507c) 所描述 (即: 703, AF 向 NWDAF 发送 Nnwdaf_AnalyticsSubscription_Subscribe (即分析订阅订阅)。实施过程:

步骤 703a, AF 向 NWDAF 发送 Nnwdaf_MLModelSplit_Request, 请求中包含和 Model ID 关联的 Analytics ID=MLModelSplit, UE 的可提供的算力、 30 可提供的内存、剩余的电量, 以及推理此模型需要的数据集的隐私等级, 以

及模型不同层的时延需求等信息，以及请求 NWDAF 帮忙发现其他可参与模型推理的 UE(s)以及其相关的可提供的算力、可提供的内存、剩余的电量等并请求 NWDAF 进行模型切分点的判断。

5 步骤 703b, AF 向 NEF 发送 Nnef_MLModelSplit_Request, 请求中包含上述步骤 703a 的信息。

步骤 703c, NEF 授权后, 向 NWDAF 发送 Nnwdaf_MLModelSplit_Request, 请求中包含上述步骤 703a 信息。

10 步骤 704, NWDAF 调用 Nnf_EventExposure_Subscribe 向 5GC NF(s) (例如, AMF/SMF) 采集 UE(s)的可提供的算力、可提供的内存、剩余的电量, (如果有)推理此模型需要的数据集的隐私等级, 以及(如果能提供)模型不同层的时延需求; 5GC NF(s) (例如, AMF/SMF) 将这些请求信息发送给终端, 终端收到这个请求, (如果通过这个请求)就准备和该 Analytics ID (Model ID) 对应的电量、内存、算力, (如果有)数据集的隐私等级等信息反馈给 NF(例如, AMF/SMF)。

15 步骤 705, 5GC NF(s)调用 Nnf_EventExposure_Notify 向 NWDAF 反馈所需数据)。

步骤 706 (即步骤 706a-706c) 如场景 13 的实施例中的步骤 507 (即步骤 507a-507c) 所描述 (即: 步骤 706, NWDAF 向 AF 发送 Nnwdaf_AnalyticsSubscription_Notify (即分析订阅通知)。实施过程:

20 步骤 706a, 通过 Nnwdaf_MLModelSplit_Response 将模型切分点 (或者模型切分比例) 等信息发送给 AF, 如表 3。

步骤 706b, NWDAF 向 NEF 发送 Nnwdaf_MLModelSplit_Response, 内容包含上述模型切分点 (或者模型切分比例) 等的信息。

25 步骤 706c, NEF 授权后, 向 NWDAF 发送 Nnef_MLModelSplit_Request response, 内容包含上述模型切分点 (或者模型切分比例) 等的信息。)

步骤 707, AF 基于接收到的信息以及自身的模型和能力, 判断相关 UE 的模型切分点(模型切分比例)。

30 步骤 708, AF 通过 Nnf_MLModelJointInference_Request response 将判断结果和相关 UE 信息以及对应模型切分点(模型切分比例)信息发送给 UE. 该信息也可以实现 UE 和 UE 之间进行联合模型推理。

场景 22、当 UE 和 AF 联合推理模型时，AF 请求 NWDAF 收集 UE 的能力，AF 基于收到的分析结果，进行模型切分点的判断，并向相关 UE 发起联合推理请求。

可选地，如果所述网络实体是应用功能 AF，所述接收用于辅助人工智能 /机器学习 AI/ML 模型切分分析的第一消息，包括：

步骤 d21、向网络数据分析功能 NWDAF 发送第十请求，所述第十请求用于请求 NWDAF 向 5GC NF(s)采集 UE(s)对应的用于分析 AI/ML 模型切分的数据，所述第十请求中携带的参数包括下述至少一项：与模型标识或模型切分标识关联的分析类型标识、接收 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足分析条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第十请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

步骤 d22、接收 NWDAF 发送的第十请求结果，所述第十请求结果包括下述至少一项：待参与 AI/ML 模型切分的 UE(s)或者 SUPI、使用 AI/ML 模型的应用的标识、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量、AI/ML 模型的大小；若第十请求中携带的参数包括推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息，则所述第十请求结果中还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

步骤 d23、将所述第十请求结果作为所述第一消息。

可选地，在确定 AI/ML 模型切分的切分结果之后，所述方法还包括：

步骤 d24、向待参与模型联合推理的 UE(s)发送第十一请求，所述第十一请求用于请求与所述待参与模型联合推理的 UE(s)执行模型联合推理操作；其中，所述第十一请求中携带的参数包括下述至少一项：模型切分点信息、所述第一消息；

步骤 d25、接收待参与模型联合推理的 UE(s)发送的第十一请求结果，所

述第十一请求结果是由待参与模型联合推理的 UE(s)根据第十一请求中携带的参数确定的，所述第十一请求结果包括接受第十一请求或不接受第十一请求。

具体地，参见图 8，图 8 为本公开实施例二提供的当网络实体是 AF 时辅助模型切分的方法的第二信令流程示意图。具体步骤为：（步骤 801a 是 AF 在受信任区域，步骤 801b、步骤 801c 是 AF 在不受信任区；步骤 804a 是 AF 在受信任区域，步骤 804b、步骤 804c 是 AF 在不受信任区。）

步骤 801（即 801a-801c）-804（即 804a-804c），AF 请求 NWDAF 收集 UE 的能力并获得分析结果如场景 11 的实施例中的步骤 301（即 301a-301c）~303、305（即 305a-305c）所描述（即：

步骤 801a，AF 向 NWDAF 发送 Nnwdaf_MLModelSplit_Request，请求中包含和 Model ID 关联的 Analytics ID（即分析类型标识）= MLModelSplit，请求 NWDAF 收集 UE（即待参与模型联合推理的 UE(s)）的可提供的算力、可提供的内存、剩余的电量，以及推理此模型需要的数据集的隐私等级，以及可选的包含模型不同层的时延需求等信息，如表 1。

步骤 801b，AF 向 NEF 发送 Nnef_MLModelSplit_Request，请求中包含上述表 1 中的信息。

步骤 801c，NEF 授权后，向 NWDAF 发送 Nnwdaf_MLModelSplit_Request，请求中包含上述表 1 中的信息。

步骤 802，NWDAF 调用 Nnf_EventExposure_Subscribe（即事件开放订阅）向 5GC NF(s）（例如，AMF/SMF）采集 UE(s)的可提供的算力、可提供的内存、剩余的电量，（如果有）推理此模型需要的数据集的隐私等级，以及（如果能提供）模型不同层的时延需求；5GC NF(s）（例如，AMF/SMF）将这些请求信息发送给终端，终端收到这个请求，（如果终端通过这个请求）就准备和该 Analytics ID（Model ID）对应的电量、内存、算力，（如果有）数据集的隐私等级等信息反馈（参见表 2）给 NF（例如，AMF/SMF）。

步骤 803，5GC NF(s)调用 Nnf_EventExposure_Notify（即事件开放通知）向 NWDAF 反馈所需数据。

步骤 804a，通过 Nnwdaf_MLModelSplit_Request Response 将模型切分点

(或者模型切分比例)等信息(如表3)发送给AF。

步骤804b, NWDAF向NEF发送Nnwdaf_MLModelSplit_Request Response, 内容包含上述模型切分点(或者模型切分比例)等的信息(参见表3)。

5 步骤804c, NEF授权后, 向NWDAF发送Nnef_MLModelSplit_Request response, 内容包含上述模型切分点(或者模型切分比例)等的信息(参见表3)。)

步骤805, AF基于接收到的信息以及自身的模型和能力, 进行模型切分点判断。

10 步骤806-808, AF和相关UE(s)建立连接并进行模型联合推理请求的响应如场景11的实施例中的步骤306-308所示(即:

步骤806, AF和相关UE(s)建立连接。

步骤807, AF向相关UE(s)发送模型联合推理请求Naf_MLModelJointInference_Request, 请求中包含模型切分点(或者模型切分比例)信息。

15 步骤808, UE(s)向AF发送模型联合推理请求响应Naf_MLModelJointInference_Request response, 表示是否接受这个模型联合推理请求。)

20 示例性地, 实施例三(网络实体为MMF(MMF可以进行模型的管理; 也可以进行模型切分), MMF进行模型切分点的判断), 下述以至少两种场景为例对辅助模型切分的方法进行详细说明。

场景31、当UE和AF联合推理模型时, AF请求新的网络实体MMF进行模型切分, MMF请求NWDAF收集UE的能力, MMF基于收到的分析结果, 进行模型切分点的判断, 并向AF反馈结果, AF向相关UE发起联合推理请求。

25 可选地, 如果所述网络实体是新的网络实体MMF, 所述接收用于辅助人工智能/机器学习AI/ML模型切分分析的第一消息, 包括:

步骤e11、直接或通过网络能力开放功能NEF接收应用功能AF发送的第十二请求, 所述第十二请求用于请求AI/ML模型切分分析; 所述第十二请求中携带的参数包括下述至少一项: 与模型标识或模型切分标识关联的分析

类型标识、接收 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足分析条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第十二请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

步骤 e12、根据所述第十二请求中携带的参数，向 NWDAF 发送第十三请求，所述第十三请求中携带的参数包括所述第十二请求中携带的参数，且所述第十三请求用于请求 NWDAF 向 5GC NF(s)采集 UE(s)的用于分析 AI/ML 模型切分的数据；

步骤 e13、接收 NWDAF 发送的第十三请求结果；其中，所述第十三请求结果包括下述至少一项：待参与 AI/ML 模型切分的 UE(s)或者 SUPI、使用 AI/ML 模型的应用的标识、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量、AI/ML 模型的大小；若第十三请求中携带的参数包括推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息，则所述第十二请求结果中还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

步骤 e14、将所述第十三请求结果作为所述第一消息。

可选地，所述确定 AI/ML 模型切分的切分结果之后，所述方法还包括：向所述 AF 发送第八消息，所述第八消息中包括所述 AI/ML 模型切分点、所述第一消息中的数据；所述第八消息用于为 AF 提供向待参与模型联合推理的 UE(s)发送第十四请求时携带的参数，所述第十四请求用于请求与待参与模型联合推理的 UE(s)执行模型联合推理操作，所述第十四请求中携带的参数包括所述第八消息。

具体地，参见图 9，图 9 为本公开实施例三提供的当网络实体是 MMF 时辅助模型切分的方法的第一信令流程示意图。具体步骤为：

步骤 901，AF 向 MMF 发送 Nmmf_MLModelSplit_Request（即机器学习模型切分请求），请求中包含的参数同场景 11 的实施例中的步骤 301 的请求

包含的参数(即:请求中包含和 Model ID 关联的 Analytics ID=MLModelSplit, 请求 NWDAF 收集 UE 的可提供的算力、可提供的内存、剩余的电量, 以及推理此模型需要的数据集的隐私等级, 以及可选的包含模型不同层的时延需求等信息, 如表 1。)

5 步骤 902-905 如场景 11 的实施例中的步骤 301c, 302, 303, 305b 所描述(即:

步骤 902, NEF 授权后, 向 NWDAF 发送 Nnwdaf_MLModelSplit_Request, 请求中包含上述表 1 中的信息。

步骤 903, NWDAF 调用 Nnf_EventExposure_Subscribe(即事件开放订阅)
10 向 5GC NF(s) (例如, AMF/SMF) 采集 UE(s) 的可提供的算力、可提供的内存、剩余的电量, (如果有)推理此模型需要的数据集的隐私等级, 以及(如果能提供)模型不同层的时延需求; 5GC NF(s) (例如, AMF/SMF) 将这些请求信息发送给终端, 终端收到这个请求, (如果终端通过这个请求)就准备和该 Analytics ID (Model ID) 对应的电量、内存、算力, (如果有)数据集
15 的隐私等级等信息反馈(参见表 2)给 NF(例如, AMF/SMF)。

步骤 904, 5GC NF(s)调用 Nnf_EventExposure_Notify 向 NWDAF 反馈所需数据。

步骤 905, NWDAF 向 NEF 发送 Nnwdaf_MLModelSplit_Request Response, 内容包含上述模型切分点(或者模型切分比例)等的信息)

20 步骤 906, MMF 基于 NWDAF 的分析输出结果 output data(参见表 3), 即基于 UE 的可提供的算力、可提供的内存、剩余的电量, 以及推理此模型需要的数据集的隐私等级, 以及模型不同层的时延需求, 选出模型切分点(或者模型切分比例)。

步骤 907, MMF 向 AF 通过 Nmmf_MLModelSplit_Request response (机器学习模型切分请求的响应)发送模型切分点(模型切分比例)等信息
25

步骤 908-9010 如场景 11 的实施例中的步骤 306-308 所描述(即:

步骤 908, AF 和相关 UE(s)建立连接。

步骤 909, AF 向相关 UE(s)发送模型联合推理请求 Naf_MLModelJointInference_Request, 请求中包含模型切分点(或者模型切分
30 比例)信息。

步骤 9010, UE(s) 向 AF 发送模型联合推理请求的响应 Naf_MLModelJointInference_Request response, 表示是否接受这个模型联合推理请求)。

场景 32、当 UE 和 AF 联合推理模型时, UE 上报自身能力给 AF 请求模型联合推理, AF 将自身以及 UE 能力传给 MMF, 请求 MMF 进行模型切分, MMF 判断如果需要其他 UE(s) 参与模型联合推理, 则 MMF 请求 NWDAF 收集其他 UE(s) 的信息, MMF 基于收集的信息进行模型切分并将结果反馈给 AF。

可选地, 如果所述网络实体是新的网络实体 MMF, 所述接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息, 包括:

步骤 e21、直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第十五请求, 所述第十五请求用于请求 AI/ML 模型切分分析, 所述第十五请求中携带的参数包括下述至少一项: 与模型标识或模型切分标识关联的分析类型标识、待参与模型联合推理的 UE(s) 可提供的算力、待参与模型联合推理的 UE(s) 可提供的内存、待参与模型联合推理的 UE(s) 剩余的电量; 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 所述第十五请求中携带的参数还包括: 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

步骤 e22、根据所述第十五请求中携带的参数, 确定 AF 与所述待参与模型联合推理的 UE(s) 是否具备完全支持执行模型联合推理的能力;

步骤 e23、若与所述待参与模型联合推理的 UE(s) 具备完全支持执行模型联合推理的能力, 则将所述第十五请求中携带的参数作为所述第一消息;

步骤 e24、若与所述待参与模型联合推理的 UE(s) 不具备完全支持执行模型联合推理的能力, 则向 NWDAF 发送第十六请求, 所述第十六请求中携带的参数包括所述第十五请求中携带的参数, 且所述第十六请求用于请求 NWDAF 通过 NF 查找其他可参与模型推理的 UE(s) 以及其他可参与模型推理的 UE(s) 可提供的算力、可提供的内存、剩余的电量, 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 所述第十六请求中携带的参数还包括: 其他可参与模型推理的 UE(s) 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

步骤 e25、接收 NWDAF 发送的第十六请求结果，并将所述第十六请求结果以及第十五请求中携带的参数作为所述第一消息；其中，所述第十六请求结果包括其他可参与模型推理的 UE(s)以及其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量，若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第十六请求结果还包括：其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息。（其中，所述第十六请求结果是由 NWDAF 通过向 5G CNF(s)采集 UE(s)对应的第四数据确定的；所述第四数据中包括下述至少一项：待参与 AI/ML 模型切分的 UE(s)或者 SUPI、待参与模型联合推理的 UE(s)（这里的待参与模型联合推理的 UE(s)为上述的其他可参与模型推理的 UE(s)）可提供的算力、待参与模型联合推理的 UE(s)（这里的待参与模型联合推理的 UE(s)为上述的其他可参与模型推理的 UE(s)）可提供的内存、待参与模型联合推理的 UE(s)（这里的待参与模型联合推理的 UE(s)为上述的其他可参与模型推理的 UE(s)）剩余的电量、AI/ML 模型的大小；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第四数据中还包括：其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；所述第十六请求结果中包括下述至少一项：第四数据、其他可参与模型推理的 UE(s)以及其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量，若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第十六请求结果还包括：其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息。）

可选地，所述确定 AI/ML 模型切分的切分结果之后，所述方法还包括：向所述 AF 发送第十五请求结果，所述第十五请求结果中包括 AI/ML 模型切分点信息；

其中，若存在其他参与联合模型推理的 UE(s)，则所述第十五请求结果中还包括其他参与联合模型推理的 UE(s)对应的模型切分点信息，并通过 AF 透传给所述待参与模型联合推理的 UE(s)。

具体地，参见图 10，图 10 为本公开实施例三提供的当网络实体是 MMF 时辅助模型切分的方法的第二信令流程示意图。具体步骤为：

步骤 1000-1001 如场景 13 的实施例中的步骤 500-501 所描述（即：

步骤 1000， AF 和 UE 建立连接。

步骤 1001， UE 发起和 AF 进行 ML 模型联合推理请求 Naf_MLModelJointInference_Request, 请求中包含和 Model ID 关联的 Analytics ID=MLModelSplit, UE 自身的可提供的算力、可提供的内存、剩余的电量等信息；如果有推理 AI/ML 模型时需要数据集的隐私等级，则请求中还包含推理 AI/ML 模型需要的数据集的隐私等级，如果没有，基于 SA3 确认是否可以设定；（如果有推理 AI/ML 模型时需要模型不同层的时延需求信息，则请求中还包含模型不同层的时延需求信息，如果没有，由 AF 提供。

10 步骤 1002， AF 向 MMF 发送模型切分请求 Nmmf_MLModelSplit_Request, 请求中包含的信息如场景 13 的实施例中的步骤 503a 所描述（即：

请求中包含和 Model ID 关联的 Analytics ID=MLModelSplit, UE 的可提供的算力、可提供的内存、剩余的电量，以及推理此模型需要的数据集的隐私等级，以及模型不同层的时延需求等信息，以及请求 NWDAF 帮忙发现其他可参与模型推理的 UE(s) 以及其相关的可提供的算力、可提供的内存、剩余的电量等。）

15 步骤 1003， MMF 判断联合推理请求，判断如果需要其他 UE(s) 参与模型联合推理，则执行步骤 1004-1007，否则跳过。

20 步骤 1004-1007 如场景 13 的实施例中的步骤 503c、504、505、507b 所描述（即：

步骤 1004， NEF 授权后，向 NWDAF 发送 Nnwdaf_MLModelSplit_Request, 请求中包含上述信息。

25 步骤 1005， NWDAF 调用 Nnf_EventExposure_Subscribe 向 5GC NF(s)（例如 AMF/SMF）采集 UE(s) 的可提供的算力、可提供的内存、剩余的电量，（如果有）推理此模型需要的数据集的隐私等级，以及（如果能提供）模型不同层的时延需求；5GC NF(s）（例如， AMF/SMF）将这些请求信息发送给终端，终端收到这个请求，（如果通过这个请求）就准备和该 Analytics ID（Model ID）对应的电量、内存、算力，（如果有）数据集的隐私等级等信息反馈给 NF（例如， AMF/SMF）。

30 步骤 1006， 5GC NF(s) 调用 Nnf_EventExposure_Notify 向 NWDAF 反馈

所需数据;

步骤 1007, NWDAF 向 NEF 发送 Nnwdaf_MLModelSplit_Request Response, 内容包含上述模型切分点 (或者模型切分比例) 等的信息)

步骤 1008, MMF 基于收集到的信息进行模型切分的判断

5 步骤 1009, MMF 将结果通过 Nmmf_MLModelSplit_Request response 发送给 AF, 响应中包含的信息如场景 11 的实施例中的步骤 305 所描述 (即: 模型切分点 (或者模型切分比例) 等信息, 如表 3)

步骤 1010 如场景 13 的实施例中的步骤 508 所描述。(即:

10 步骤 1010, AF 把模型联合推理请求的响应 Naf_MLModelJointInference_Request response 发送给 UE, 响应中包含模型切分点 (或者模型切分比例) 以及其他可参与模型推理的 UE(s) 以及其相关的模型切分点 (或者模型切分比例) 信息)。

示例性地, 实施例四 (网络实体为 PCF, PCF 进行模型切分点的判断), 下述以至少两种场景为例对辅助模型切分的方法进行详细说明。

15 场景 41、当 UE 和 AF 联合推理模型时, AF 请求 PCF 进行模型切分的策略判断, PCF 请求 NWDAF 收集 UE 的能力, PCF 基于收到的分析结果, 进行模型切分点的判断, 并向 AF 反馈结果, AF 向相关 UE 发起联合推理请求。

可选地, 如果所述网络实体是策略控制功能 PCF, 所述接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息, 包括:

20 步骤 f11、直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第十七请求, 所述第十七请求用于请求 AI/ML 模型切分分析; 所述第十七请求中携带的参数包括下述至少一项: 与模型标识或模型切分标识关联的分析类型标识、接收 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足分析条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE(s) 可提供的算力、待参与模型联合推理的 UE(s) 可提供的内存、待参与模型联合推理的 UE(s) 剩余的电量; 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 所述第十七请求中携带的参数还包括: 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

30 步骤 f12、根据所述第十七请求中携带的参数, 向网络数据分析功能

NWDAF 发送第十八请求, 所述第十八请求中携带的参数包括所述第十七请求中携带的参数, 且所述第十八请求用于请求 NWDAF 向 5GC NF(s) 采集 UE(s) 对应的第五数据; 所述第五数据中包括下述至少一项: 待参与 AI/ML 模型切分的 UE(s) 或者 SUPI、待参与模型联合推理的 UE(s) 可提供的算力、待参与模型联合推理的 UE(s) 可提供的内存、待参与模型联合推理的 UE(s) 剩余的电量、AI/ML 模型的大小; 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 所述第五数据中还包括: 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

步骤 f13、接收 NWDAF 发送的第十八请求结果; 其中, 所述第十八请求结果包括所述第五数据;

步骤 f14、将所述第十八请求结果作为所述第一消息。

可选地, 所述确定 AI/ML 模型切分的切分结果之后, 所述方法还包括: 向所述 AF 发送第九消息, 所述第九消息中包括所述 AI/ML 模型切分点、所述第一消息的数据; 所述第九消息用于为 AF 提供向待参与模型联合推理的 UE(s) 发送第十九请求时携带的参数, 所述第十九请求用于请求与待参与模型联合推理的 UE(s) 执行模型联合推理操作, 所述第十九请求中携带的参数包括所述第九消息。

具体地, 参见图 11, 图 11 为本公开实施例四提供的当网络实体是 PCF 时辅助模型切分的方法的第一信令流程示意图。具体步骤为:

步骤 1101, AF 向 PCF 发送 Npcf_MLModelSplit_Request (即机器学习模型切分请求), 请求中包含的参数同场景 11 的实施例中的步骤 301 的请求 (即: 请求中包含和 Model ID 关联的 Analytics ID= MLModelSplit, 请求 NWDAF 收集 UE 的可提供的算力、可提供的内存、剩余的电量, 以及推理此模型需要的数据集的隐私等级, 以及可选的包含模型不同层的时延需求等信息, 如表 1。)

步骤 1102-1105 如场景 11 的实施例中的步骤 301c, 302, 303, 305b 所描述 (即:

步骤 1102, NEF 授权后, 向 NWDAF 发送 Nnwdaf_MLModelSplit_Request, 请求中包含上述表 1 中的信息。

步骤 1103, NWDAF 调用 Nnf_EventExposure_Subscribe (即事件开放订

阅)向 5GC NF(s) (例如, AMF/SMF) 采集 UE(s)的可提供的算力、可提供的内存、剩余的电量, (如果有)推理此模型需要的数据集的隐私等级, 以及(如果能提供)模型不同层的时延需求; 5GC NF(s) (例如, AMF/SMF) 将这些请求信息发送给终端, 终端收到这个请求, (如果终端通过这个请求)就准备和该 Analytics ID (Model ID) 对应的电量、内存、算力, (如果有)数据集的隐私等级等信息反馈 (参见表 2) 给 NF(例如, AMF/SMF)。

步骤 1104, 5GC NF(s)调用 Nnf_EventExposure_Notify 向 NWDAF 反馈所需数据。

步骤 1105, NWDAF 向 NEF 发送 Nnwdaf_MLModelSplit_Request Response, 内容包含上述模型切分点 (或者模型切分比例) 等的信息)。

步骤 1106, PCF 基于 NWDAF 的分析输出结果 output data, 即基于 UE 的可提供的算力、可提供的内存、剩余的电量, 以及推理此模型需要的数据集的隐私等级, 以及模型不同层的时延需求, 选出模型切分点 (或者模型切分比例)。

例如, UE 的可提供的算力和内存较少、剩余的电量较低时, 只能做较少层的推理, 如果推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 且隐私等级需求较高和/或模型不同层的时延需求较低时做较少层的推理; 如果 UE 的可提供的算力和内存充足、剩余的电量较高时, 可以做较多层的推理, 如果推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 且隐私等级需求较低和/或模型不同层的时延需求较高时做较多层的推理; 例如, 假设对于一个 8 层的 AI/ML 模型, UE1 可以做两层的推理 (比如, UE1 做前两层的推理)、UE2 可以做三层的推理 (比如, UE2 做中间三层的推理), AF 做最后三层的推理, 即 UE1 对应模型切分点 1, UE2 对应模型切分点 2。

步骤 1107, PCF 向 AF 通过 Npcf_MLModelSplit_Request response (即机器学习模型切分请求的响应) 发送模型切分点(模型切分比例)等信息。

步骤 1108-1110 如场景 11 的实施例中的步骤 306-308 所描述 (即:

步骤 1108, AF 和相关 UE(s)建立连接。

步骤 1109, AF 向相关 UE(s) 发送模型联合推理请求 Naf_MLModelJointInference_Request, 请求中包含模型切分点 (或者模型切分

比例)信息。

步骤 1110, UE(s)向 AF 发送模型联合推理请求的响应 Naf_MLModelJointInference_Request response, 表示是否接受这个模型联合推理请求)。

- 5 场景 42、当 UE 和 AF 联合推理模型时, UE 上报自身能力给 AF 请求模型联合推理, AF 将自身以及 UE 能力发给 PCF, 请求 PCF 进行模型切分, PCF 判断如果需要其他 UE(s)参与模型联合推理, 则 PCF 请求 NWDAF 收集其他 UE(s)的信息, PCF 基于收集的信息进行模型切分并将结果反馈给 AF。

可选地, 如果所述网络实体是策略控制功能 PCF, 所述接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息, 可以包括以下步骤:

步骤 f21、直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第二十请求, 所述第二十请求用于请求 AI/ML 模型切分分析, 所述第二十请求中携带的参数包括下述至少一项: 与模型标识或模型切分标识关联的分析类型标识、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量; 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 所述第二十请求中携带的参数还包括: 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

步骤 f22、根据所述第二十请求中携带的参数, 确定所述 AF 以及待参与模型联合推理的 UE 是否具备完全支持执行模型联合推理的能力;

步骤 f23、若所述 AF 以及待参与模型联合推理的 UE(s)具备完全支持执行模型联合推理的能力, 则将所述第二十请求中携带的参数作为所述第一消息;

步骤 f24、若所述 AF 以及待参与模型联合推理的 UE(s)不具备完全支持执行模型联合推理的能力, 则向 NWDAF 发送第二十一请求, 所述第二十一请求中携带的参数包括所述第二十请求中携带的参数, 且所述第二十一请求用于请求 NWDAF 通过 NF 查找其他可参与模型推理的 UE(s)以及其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量, 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 所述第二十一请求中携带的参数还包括: 其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数

据集的隐私等级和/或模型不同层的时延需求信息;

步骤 f25、接收 NWDAF 发送的第二十一请求结果, 并将所述第二十一请求结果以及所述第二十请求中携带的参数作为所述第一消息; 其中, 所述第二十一请求结果包括: 其他可参与模型推理的 UE(s) 以及其他可参与模型推理的 UE(s) 可提供的算力、可提供的内存、剩余的电量, 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 所述第二十一请求结果还包括: 其他可参与模型推理的 UE(s) 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息; (所述第二十一请求结果是由 NWDAF 通过向 5GC NF(s) 采集 UE(s) 对应的第六数据确定的; 所述第六数据中包括下述至少一项: 待参与 AI/ML 模型切分的 UE(s) 或者 SUPI、待参与模型联合推理的 UE(s) (这里的待参与模型联合推理的 UE(s) 为上述的其他可参与模型推理的 UE(s)) 可提供的算力、待参与模型联合推理的 UE(s) (这里的待参与模型联合推理的 UE(s) 为上述的其他可参与模型推理的 UE(s)) 可提供的内存、待参与模型联合推理的 UE(s) (这里的待参与模型联合推理的 UE(s) 为上述的其他可参与模型推理的 UE(s)) 剩余的电量、AI/ML 模型的大小; 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 所述第六数据中还包括: 其他可参与模型推理的 UE(s) 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息; 所述第二十一请求结果中包括下述至少一项: 第六数据、其他可参与模型推理的 UE(s) 以及其他可参与模型推理的 UE(s) 可提供的算力、可提供的内存、剩余的电量, 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 所述第二十一请求结果还包括: 其他可参与模型推理的 UE(s) 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息。)

步骤 f26、若第二十请求中携带的参数中不包含模型不同层的时延需求信息, 则模型不同层的时延需求信息由 AF 提供。

可选地, 所述确定 AI/ML 模型切分的切分结果之后, 所述方法还包括: 向所述 AF 发送第二十请求结果, 所述第二十请求结果中包括 AI/ML 模型切分点信息;

其中, 若存在其他参与联合模型推理的 UE(s), 则所述第二十请求结果中还包括其他参与联合模型推理的 UE(s) 对应的模型切分点信息, 并通过 AF 透

传给所述 UE。

具体地，参见图 12，图 12 为本公开实施例四提供的当网络实体是 PCF 时辅助模型切分的方法的第二信令流程示意图。具体步骤为：

步骤 1200-1201 如场景 13 的实施例中的步骤 500-501 所描述（即：

5 步骤 1200，AF 和 UE 建立连接

步骤 1201，UE 发起和 AF 进行 ML 模型联合推理请求 Naf_MLModelJointInference_Request，请求中包含和 Model ID 关联的 Analytics ID=MLModelSplit，UE 自身的可提供的算力、可提供的内存、剩余的电量等信息；如果有推理 AI/ML 模型时需要数据集的隐私等级，则请求中还包含推理 AI/ML 模型需要的数据集的隐私等级，如果没有，基于 SA3 确认是否可以设定；（如果有推理 AI/ML 模型时需要模型不同层的时延需求信息，则请求中还包含模型不同层的时延需求信息，如果没有，由 AF 提供。

步骤 1202，AF 向 PCF 发送模型切分请求 Npcf_MLModelSplit_Request，请求中包含的信息如场景 13 的实施例中的步骤 503a 所描述（即请求中包含和 Model ID 关联的 Analytics ID=MLModelSplit，UE 的可提供的算力、可提供的内存、剩余的电量，以及推理此模型需要的数据集的隐私等级，以及模型不同层的时延需求等信息，以及请求 NWDAF 帮忙发现其他可参与模型推理的 UE(s)以及其相关的可提供的算力、可提供的内存、剩余的电量等）

步骤 1203，基于接收到的信息，例如，UE、AF 报告的自身能力等，PCF 判断联合推理请求，判断如果需要其他 UE(s)参与模型联合推理，则执行步骤 1204-1207，否则跳过。

步骤 1204-1207 如场景 13 的实施例中的步骤 503c、504、505、507b 所描述（即：

步骤 1204，NEF 授权后，向 NWDAF 发送 Nnwdaf_MLModelSplit_Request，请求中包含上述信息。

步骤 1205，NWDAF 调用 Nnf_EventExposure_Subscribe 向 5GC NF(s)（例如，AMF/SMF）采集 UE(s)的可提供的算力、可提供的内存、剩余的电量，（如果有）推理此模型需要的数据集的隐私等级，以及（如果能提供）模型不同层的时延需求；5GC NF(s）（例如，AMF/SMF）将这些请求信息发送给终端，终端收到这个请求，（如果通过这个请求）就准备和该 Analytics ID（Model

ID)对应的电量、内存、算力, (如果有)数据集的隐私等级等信息反馈给NF(例如, AMF/SMF)。

步骤 1206, 5GC NF(s)调用 Nnf_EventExposure_Notify 向 NWDAF 反馈所需数据。

5 步骤 1207, NWDAF 向 NEF 发送 Nnwdaf_MLModelSplit_Request Response, 内容包含上述模型切分点 (或者模型切分比例) 等的信息。)

步骤 1208, PCF 基于收集到的信息进行模型切分的判断。

步骤 1209, PCF 将结果通过 Nmmf_MLModelSplit_Request response 发送给 AF, 响应中包含的信息如场景 11 的实施例中的步骤 305 所描述 (即: 模型切分点 (或者模型切分比例) 等信息, 如表 3。)

步骤 1210 如场景 13 的实施例中的步骤 508 所描述。(即:

步骤 1210, AF 把模型联合推理请求的响应 Naf_MLModelJointInference_Request response 发送给 UE, 响应中包含模型切分点 (或者模型切分比例) 以及其他可参与模型推理的 UE(s) 以及其相关的模型切分点 (或者模型切分比例) 信息)。

本公开实施例, 通过 PCF 接收 AF 发送的模型切分请求, 以及请求中包含的参数; PCF 进行模型切分点的判断, 如果需要, 从 NWDAF 请求分析结果。或者, 通过 AF 接收 UE 发送的模型切分请求, 以及请求中包含的参数; AF 基于自身信息或者从 NWDAF 请求得到的分析结果进行模型切分点的判断; AF 向 UE 发送分析和判断结果。或者, 通过引入新的模型管理网络实体 MMF, MMF 接收 AF 发送的模型切分请求, 以及请求中包含的参数 MMF 进行模型切分点的判断, 如果需要, 从 NWDAF 请求分析结果。或者, NWDAF 接收 AF/终端/PCF/新网络实体等发送的模型切分请求, 以及请求中包含的参数; NWDAF 从 5GC NF(s)采集的输入数据, 用于模型切分的分析; (如果其他不能做, NWDAF 进行分析和模型切分判断), 向 AF/终端/PCF/新网络实体等发送分析和判断结果。因此, 网络实体基于终端的电量、可提供的内存、可提供的算力等信息, 以及关于此模型的时延需求以及关于此模型需要的推理数据集的隐私等级等, 选择出模型切分点, 网络实体将模型切分点信息发送给参与模型联合推理的网络实体或者终端。基于终端能力实现模型切分推理, 从而有益于保护终端隐私, 优化网络资源。

图 13 为本公开实施例提供的辅助模型切分的方法的第二流程示意图，如图 13 所示，本实施例提供的辅助模型切分的方法的执行主体为用户设备（或终端）UE，则本公开实施例提供的辅助模型切分的方法包括以下步骤：

步骤 201、UE 根据自身能力信息，确定 AI/ML 模型切分点信息。

5 可选地，根据自身能力信息，确定 AI/ML 模型切分点信息，可以包括以下步骤：

步骤 g1、根据自身能力信息中的可提供的算力、可提供的内存以及剩余的电量，确定模型切分点信息；

10 步骤 g2、若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，则根据所述自身能力信息中的可提供的算力、可提供的内存、剩余的电量、推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息，确定模型切分点信息。

步骤 202、UE 将所述 AI/ML 模型切分点信息作为 AI/ML 模型切分的切分结果。

15 具体地，终端可以基于自身可提供的电量、可提供的内存、可提供的算力等信息，以及关于此模型的时延需求以及关于此模型需要的推理数据集的隐私等级等，选择出模型切分点（即模型切分点信息或 AI/ML 模型切分点信息），然后终端将模型切分点信息发送给参与模型联合推理的网络实体或者终端。

20 可选地，确定模型切分点信息，可以包括以下步骤：

将所有确定参与执行模型联合推理的 UE 中的任一 UE 作为目标 UE，针对每个所述目标 UE 执行下述步骤：

25 步骤 h1、若所述目标 UE 可提供的算力低于第一预设算力阈值或所述目标 UE 可提供的内存低于第一预设内存阈值或所述目标 UE 剩余的电量低于第一预设电量阈值时，确定所述目标 UE 执行第一预设数目层的推理；

30 步骤 h2、若所述目标 UE 可提供的算力高于第一预设算力阈值并低于第二预设算力阈值，和所述目标 UE 可提供的内存高于第一预设内存阈值并低于第二预设内存阈值，以及所述目标 UE 剩余的电量高于第一预设电量阈值并低于第二预设电量阈值时，确定所述目标 UE 执行第二预设数目的推理，第一预设数目层小于第二预设数目层；

步骤 h3、若所述目标 UE 可提供的算力高于第二预设算力阈值并低于第三预设算力阈值，和所述目标 UE 可提供的内存高于第二预设内存阈值并低于第三预设内存阈值，以及所述目标 UE 剩余的电量高于第二预设电量阈值并低于第三预设电量阈值时，确定所述目标 UE 执行第三预设数目层的推理，
5 第二预设数目层小于第三预设数目层；以此类推，直至确定所述目标 UE 执行第 N+1 预设数目层的推理，第 N 预设数目层小于第 N+1 预设数目层；

步骤 h4、若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，且若推理 AI/ML 模型需要的数据集的隐私等级低于第一预设隐私等级或模型不同层的时延需求信息低于第一预设时延，则所述目标 UE 执行所述
10 第一预设数目层的推理；

步骤 h5、若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，且若推理 AI/ML 模型需要的数据集的隐私等级高于第一预设隐私等级并低于第二预设隐私等级，和模型不同层的时延需求信息高于第一预设时延并低于第二预设时延时，确定所述目标 UE 执行第二预设数目层的推理；

15 步骤 h6、若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，且若推理 AI/ML 模型需要的数据集的隐私等级高于第二预设隐私等级并低于第三预设隐私等级，和模型不同层的时延需求信息高于第二预设时延并低于第三预设时延时，确定所述目标 UE 执行第三预设数目层的推理；
以此类推，直至确定所述目标 UE 执行第 N+1 预设数目层的推理，第 N 预设
20 数目层小于第 N+1 预设数目层；

步骤 h7、根据预设数目层确定 AI/ML 模型切分点信息，所述 AI/ML 模型切分点信息用于表示 AI/ML 模型切分比例，所述预设数目层包括第 N 预设数目层，N 大于或等于一。

具体地，参与模型联合推理的每个 UE 都需要确定其推理的层数或部分，
25 针对每个 UE（即目标 UE）可以通过以下过程实现切分点的确定：

目标 UE 的可提供的算力和内存较少、剩余的电量较低时，只能做较少层的推理，如果推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，且隐私等级需求较高和/或模型不同层的时延需求较低时做较少层的推理；如果 UE 的可提供的算力和内存充足、剩余的电量较高时，可以做较多
30 层的推理，如果推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的

时延, 且隐私等级需求较低和/或模型不同层的时延需求较高时做较多层的推理。例如, 假设对于一个 8 层的 AI/ML 模型, UE1 可以做两层的推理(比如, UE1 做前两层的推理)、UE2 可以做三层的推理(比如, UE2 做中间三层的推理), AF 做最后三层的推理, 即 UE1 对应模型切分点 1, UE2 对应模型切分点 2。

本实施例中, 根据自身能力信息中的可提供的算力、可提供的内存以及剩余的电量, 确定模型切分点信息; 如果推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 还可以结合自身能力信息中的推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息来确定切分 AI/ML 模型的模型切分点信息, 作为切分结果, 用以实现网络实体或待参与模型联合推理的 UE(s) 基于切分结果对模型进行联合推理操作, 因此, 实现了基于终端(即 UE)能力对模型切分的分析, 进而有效地实现对终端隐私的保护以及网络资源的优化。

示例性地, 实施例五(UE 进行模型切分点的判断), 下述以至少两种场景为例对辅助模型切分的方法进行详细说明。

场景 51、当 UE 和 AF 联合推理模型时, UE 基于自身能力进行模型切分点的判断, 上报模型切分点(模型切分比例)给 AF, 并进行联合推理交互。

可选地, 所述确定 AI/ML 模型切分的切分结果之后, 所述方法还可以包括以下步骤:

步骤 i1、向应用功能 AF 发送第一请求, 所述第一请求用于请求与所述 AF 执行模型联合推理操作; 其中, 所述第一请求中携带的参数包括下述至少一项 AI/ML 模型切分点、与模型标识或模型切分标识关联的分析类型标识、接收 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足分析条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE 可提供的算力、待参与模型联合推理的 UE 可提供的内存、待参与模型联合推理的 UE 剩余的电量; 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 所述第一请求中携带的参数还包括: 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

步骤 i2、接收 AF 发送的第一请求结果, 所述第一请求结果是由 AF 根据

第一请求中携带的参数确定的，所述第一请求结果包括接受第一请求或不接受第一请求；

其中，若第一请求中携带的参数中不包含模型不同层的时延需求信息，则模型不同层的时延需求信息由 AF 提供。

5 具体地，实施过程为：

步骤 1401，在场景 12 的实施例中的步骤 401（即：

UE 发起和 AF 进行 ML 模型联合推理请求 Naf_MLModelJointInference_Request，请求中包含和 Model ID 关联的 Analytics ID=MLModelSplit，UE 自身的可提供的算力、可提供的内存、剩余的电量等信息；如果有推理 AI/ML 模型时需要数据集的隐私等级，则请求中还包含推理 AI/ML 模型需要的数据集的隐私等级，如果没有，基于 SA3 确认是否可以设定；（如果有推理 AI/ML 模型时需要模型不同层的时延需求信息，则请求中还包含模型不同层的时延需求信息，如果没有，由 AF 提供。请求中还增加模型切分点(模型切分比例)信息上报给 AF。

15 步骤 1402，如场景 22 的实施例中的步骤 806-808 所示，UE 和 AF 建立联合模型推理。（即：

AF 和相关 UE(s)建立连接。

AF 向相关 UE(s) 发送模型联合推理请求 Naf_MLModelJointInference_Request，请求中包含模型切分点（或者模型切分比例）信息。

20 UE(s) 向 AF 发送模型联合推理请求的响应 Naf_MLModelJointInference_Request response，表示是否接受这个模型联合推理请求。）

场景 52、当 UE 和 NWDAF 联合推理模型时，UE 基于自身能力进行模型切分点的判断，上报 NWDAF，并进行联合推理交互。

可选地，所述确定 AI/ML 模型切分的切分结果之后，所述方法还可以包括以下步骤：

步骤 j1、通过 AF 向网络数据分析功能 NWDAF 发送第二请求，所述第二请求用于请求与所述 NWDAF 执行模型联合推理操作；其中，所述第二请

求中携带的参数包括下述至少一项 AI/ML 模型切分点、与模型标识或模型切分标识关联的分析类型标识、接收 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足分析条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE 可提供的算力、待参与模型联合推理的 UE 可提供的内存、待参与模型联合推理的 UE 剩余的电量;若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,所述第二请求中携带的参数还包括:推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

步骤 j2、接收 NWDAF 发送的第二请求结果,所述第二请求结果是由 NWDAF 根据第一请求中携带的参数确定的,所述第二请求结果包括接受第一请求或不接受第一请求;

其中,若第二请求中携带的参数中不包含模型不同层的时延需求信息,则模型不同层的时延需求信息由 AF 提供。

具体地,实施过程为:

步骤 1501,在场景 12 的实施例中的步骤 401 (即:

UE 发起和 AF 进行 ML 模型联合推理请求 Naf_MLModelJointInference_Request,请求中包含和 Model ID 关联的 Analytics ID=MLModelSplit,UE 自身的可提供的算力、可提供的内存、剩余的电量等信息;如果有推理 AI/ML 模型时需要数据集的隐私等级,则请求中还包含推理 AI/ML 模型需要的数据集的隐私等级,如果没有,基于 SA3 确认是否可以设定;(如果有推理 AI/ML 模型时需要模型不同层的时延需求信息,则请求中还包含模型不同层的时延需求信息,如果没有,由 AF 提供。请求中还增加模型切分点(模型切分比例)信息上报给 AF。

步骤 1502,AF 通过场景 13 的实施例中的步骤 503 (即:

AF 向 NWDAF 发送 Nnwdaf_MLModelSplit_Request,请求中包含和 Model ID 关联的 Analytics ID=MLModelSplit,UE 的可提供的算力、可提供的内存、剩余的电量,以及推理此模型需要的数据集的隐私等级,以及模型不同层的时延需求等信息,以及请求 NWDAF 帮忙发现其他可参与模型推理的 UE(s)以及其相关的可提供的算力、可提供的内存、剩余的电量等(若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,还请求提供其

他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息) 并请求 NWDAF 进行模型切分点的判断;

AF 向 NEF 发送 Nnef_MLModelSplit_Request, 请求中包含上述信息 NEF 授权后, 向 NWDAF 发送 Nnwdaf_MLModelSplit_Request, 请求中包含上述
5 信息) 将信息发送给 NWDAF)。

步骤 1503, NWDAF 和 UE 建立联合模型推理。

本公开实施例, 通过 UE 发送模型联合推理请求, 以及请求中包含的参数; UE 接收从 NWDAF/AF 获得模型联合推理响应, 以及响应中包含的参数。因此, 终端基于可提供的电量、可提供的内存、可提供的算力等信息, 以及
10 关于此模型的时延需求以及关于此模型需要的推理数据集的隐私等级等, 选择出模型切分点, 终端将模型切分点信息发送给参与模型联合推理的网络实体或者终端。基于终端能力实现模型切分推理, 从而有益于保护终端隐私, 优化网络资源。

图 14 为本公开实施例提供的辅助模型切分的装置的结构示意图, 如图
15 14 所示, 本实施例提供的辅助模型切分的装置应用于网络实体。则本实施例提供的辅助模型切分的装置包括: 收发机 1400, 用于在处理器 1410 的控制下接收和发送数据。

其中, 在图 14 中, 总线架构可以包括任意数量的互联的总线和桥, 具体由处理器 1410 代表的一个或多个处理器和存储器 1420 代表的存储器的各种
20 电路链接在一起。总线架构还可以将诸如外围设备、稳压器和功率管理电路等之类的各种其他电路链接在一起, 这些都是本领域所公知的, 因此, 本文不再对其进行进一步描述。总线接口提供接口。收发机 1400 可以是多个元件, 即包括发送机和接收机, 提供用于在传输介质上与各种其他装置通信的单元, 这些传输介质包括无线信道、有线信道、光缆等传输介质。处理器 1410 负责
25 管理总线架构和通常的处理, 存储器 1420 可以存储处理器 140 在执行操作时所使用的数据。

处理器 1410 可以是中央处理器 (CPU)、专用集成电路 (Application Specific Integrated Circuit, ASIC)、现场可编程门阵列 (Field-Programmable Gate Array, FPGA) 或复杂可编程逻辑器件 (Comple6 Programmable Logic
30 Device, CPLD), 处理器也可以采用多核架构。

本实施例中，存储器 1420，用于存储计算机程序；收发机 1400，用于在处理器 1410 的控制下收发数据；处理器 1410，用于读取存储器中的计算机程序并执行以下操作：

接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息；

5 根据所述第一消息，确定 AI/ML 模型切分的切分结果。

可选地，所述第一消息中包括下述至少一项：待参与 AI/ML 模型切分的 UE(s)或者用户永久标识 SUPI、使用 AI/ML 模型的应用的标识、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量、AI/ML 模型的大小；

10 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第一消息中还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息。

可选地，处理器 1410，用于根据所述第一消息，确定 AI/ML 模型切分的切分结果时，具体包括：

15 根据所述第一消息中的待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量，确定所述待参与模型联合推理的 UE(s)对应的模型切分点信息；

20 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，则根据所述第一消息中的待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量、推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息，确定所述待参与模型联合推理的 UE(s)对应的模型切分点信息；

将所述模型切分点信息作为所述切分结果。

25 可选地，处理器 1410，用于确定所述待参与模型联合推理的 UE(s)对应的模型切分点信息时，具体包括：

将所有确定参与执行模型联合推理的 UE 中的任一 UE 作为目标 UE，针对每个所述目标 UE 执行下述步骤：

30 若所述目标 UE 可提供的算力低于第一预设算力阈值或所述目标 UE 可提供的内存低于第一预设内存阈值或所述目标 UE 剩余的电量低于第一预设电量阈值时，确定所述目标 UE 执行第一预设数目层的推理；

若所述目标 UE 可提供的算力高于第一预设算力阈值并低于第二预设算力阈值，和所述目标 UE 可提供的内存高于第一预设内存阈值并低于第二预设内存阈值，以及所述目标 UE 剩余的电量高于第一预设电量阈值并低于第二预设电量阈值时，确定所述目标 UE 执行第二预设数目的推理，第一预设数目层小于第二预设数目层；

若所述目标 UE 可提供的算力高于第二预设算力阈值并低于第三预设算力阈值，和所述目标 UE 可提供的内存高于第二预设内存阈值并低于第三预设内存阈值，以及所述目标 UE 剩余的电量高于第二预设电量阈值并低于第三预设电量阈值时，确定所述目标 UE 执行第三预设数目层的推理，第二预设数目层小于第三预设数目层；以此类推，直至确定所述目标 UE 执行第 N+1 预设数目层的推理，第 N 预设数目层小于第 N+1 预设数目层；

若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，且若推理 AI/ML 模型需要的数据集的隐私等级低于第一预设隐私等级或模型不同层的时延需求信息低于第一预设时延，则所述目标 UE 执行所述第一预设数目层的推理；

若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，且若推理 AI/ML 模型需要的数据集的隐私等级高于第一预设隐私等级并低于第二预设隐私等级，和模型不同层的时延需求信息高于第一预设时延并低于第二预设时延时，确定所述目标 UE 执行第二预设数目层的推理；

若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，且若推理 AI/ML 模型需要的数据集的隐私等级高于第二预设隐私等级并低于第三预设隐私等级，和模型不同层的时延需求信息高于第二预设时延并低于第三预设时延时，确定所述目标 UE 执行第三预设数目层的推理；以此类推，直至确定所述目标 UE 执行第 N+1 预设数目层的推理，第 N 预设数目层小于第 N+1 预设数目层；

根据预设数目层确定 AI/ML 模型切分点信息，所述 AI/ML 模型切分点信息用于表示 AI/ML 模型切分比例，所述预设数目层包括第 N 预设数目层，N 大于或等于一。

可选地，处理器 1410，用于在所述网络实体是网络数据分析功能 NWDAF 时，且接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息时，

具体包括:

直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第一请求, 所述第一请求用于请求对 AI/ML 模型切分进行分析; 其中, 所述第一请求中携带的参数包括下述至少一项: 与模型切分关联的分析类型标识、接受 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足切分条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE(s) 可提供的算力、待参与模型联合推理的 UE(s) 可提供的内存、待参与模型联合推理的 UE(s) 剩余的电量; 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 所述第一请求中携带的参数还包括: 推理 AI/ML 模型需要
5 的数据集的隐私等级和/或模型不同层的时延需求信息;

根据所述第一请求中携带的参数, 向 5GC NF(s) 发送第二消息, 所述第二消息用于请求 5GC NF(s) 采集 UE(s) 对应的第一数据, 所述第一数据中包括下述至少一项: 待参与 AI/ML 模型切分的 UE(s) 以及 SUPI、待参与模型联合推理的 UE(s) 可提供的算力、待参与模型联合推理的 UE(s) 可提供的内存、待参与模型联合推理的 UE(s) 剩余的电量、AI/ML 模型的大小; 若第一请求中携带的参数包括推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息, 则所述第一数据中还包括: 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;
15

接收 5GC NF(s) 发送的所述第一数据, 并将所述第一数据作为所述第一消息; 所述第一数据是由待参与模型联合推理的 UE(s) 同意所述第二消息的请求后提供的。
20

可选地, 处理器 1410, 还用于:

确定 AI/ML 模型切分的切分结果之后, 向所述 AF 发送第三消息, 所述第三消息中包括 AI/ML 模型切分点信息以及所述第一数据; 所述第三消息用于为 AF 提供向待参与模型联合推理的 UE(s) 发送第二请求时携带的参数, 所述第二请求用于请求与待参与模型联合推理的 UE(s) 执行模型联合推理操作, 所述第二请求中携带的参数包括所述第三消息。
25

可选地, 处理器 1410, 用于在所述网络实体是网络数据分析功能 NWDAF 时, 且接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息时, 具体包括:
30

直接或通过网络能力开放功能 NEF 接收 AF 发送的第三请求；所述第三请求是由 AF 确定与待参与模型联合推理的 UE(s) 具备完全支持执行模型联合推理的能力后确定的，其中，是否具备完全支持执行模型联合推理的能力是由 AF 根据接收到的待参与模型联合推理的 UE(s) 发送的第四请求中携带的参数确定的，所述第四请求用于待参与模型联合推理的 UE(s) 请求与 AF 进行模型联合推理，所述第四请求中携带的参数包括下述至少一项：与模型标识或模型切分标识关联的分析类型标识、待参与模型联合推理的 UE(s) 可提供的算力、待参与模型联合推理的 UE(s) 可提供的内存、待参与模型联合推理的 UE(s) 剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第四请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息，所述第三请求中携带的参数包括所述第四请求中携带的参数；将所述第三请求中携带的参数作为所述第一消息；

其中，若第四请求中携带的参数中不包含模型不同层的时延需求信息，则模型不同层的时延需求信息由 AF 提供。

可选地，处理器 1410，还用于：

确定 AI/ML 模型切分的切分结果之后，向所述 AF 发送第四消息，所述第四消息中包括所述 AI/ML 模型切分点、所述第一消息中的数据；所述第四消息用于为 AF 提供向待参与模型联合推理的 UE(s) 发送的第四请求结果，所述第四请求结果中包括所述 AI/ML 模型切分点信息。

可选地，处理器 1410，用于在所述网络实体是网络数据分析功能 NWDAF 时，且接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息时，具体包括：

直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第五请求，所述第五请求是由 AF 确定与待参与模型联合推理的 UE(s) 不具备完全支持执行模型联合推理的能力后确定的，其中，是否具备完全支持执行模型联合推理的能力是由 AF 根据接收到的待参与模型联合推理的 UE(s) 发送的第六请求确定的，所述第五请求用于请求对 AI/ML 模型切分进行分析以及请求查找其他可参与模型推理的 UE(s) 并提供其他可参与模型推理的 UE(s) 可提供的算力、可提供的内存、剩余的电量；其中，所述第五请求中携带的参数包括下述至

少一项：与模型切分关联的分析类型标识、接受 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足切分条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第五请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；所述第六请求用于待参与模型联合推理的 UE(s)请求与 AF 进行模型联合推理，所述第六请求中携带的参数包括所述第五请求中携带的参数；

根据所述第五请求中携带的参数，向 5GC NF(s)发送第五消息，所述第五消息用于请求 5GC NF(s)采集 UE(s)对应的第二数据，所述第二数据中包括下述至少一项：待参与 AI/ML 模型切分的 UE(s)或者 SUPI、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量、AI/ML 模型的大小；若第五请求中携带的参数包括推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息，则所述第二数据中还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

接收 5GC NF(s)发送的所述第二数据，并将所述第二数据作为所述第一消息；所述第二数据是由待参与模型联合推理的 UE(s)同意所述第五消息的请求后提供的；

其中，若第六请求中携带的参数中不包含模型不同层的时延需求信息，则模型不同层的时延需求信息由 AF 提供。

可选地，处理器 1410，还用于：

确定 AI/ML 模型切分的切分结果之后，向所述 AF 发送第六消息，所述第六消息中至少包括所述 AI/ML 模型切分点信息以及所述第二数据；所述第六消息用于为 AF 提供向待参与模型联合推理的 UE(s)发送的第六请求结果，所述第六请求结果中包括所述待参与模型联合推理的 UE(s)对应的所述 AI/ML 模型切分点信息、所述其他可参与模型推理的 UE(s)以及对应的所述 AI/ML 模型切分点信息。

可选地，处理器 1410，用于在所述网络实体是网络数据分析功能 NWDAF

时，且接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息时，具体包括：

接收待参与模型联合推理的 UE(s)发送的第七请求，所述第七请求用于请求与 NWDAF 执行模型联合推理操作；所述第七请求中携带的参数包括下述至少一项：待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第七请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；根据所述第七请求中携带的参数，确定 AF 与所述待参与模型联合推理的 UE(s)是否具备完全支持执行模型联合推理的能力；

若与所述待参与模型联合推理的 UE(s)具备完全支持执行模型联合推理的能力，则将所述第七请求中携带的参数作为所述第一消息；

若与所述待参与模型联合推理的 UE(s)不具备完全支持执行模型联合推理的能力，则请求 NF 查找其他可参与模型推理的 UE(s)并提供其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量，且若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，则请求提供其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

接收 NF 发送的查找结果，所述查找结果包括其他可参与模型推理的 UE(s)以及其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量，若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，则查找结果还包括：其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

将所述第七请求中携带的参数以及所述查找结果作为所述第一消息

其中，若第七请求中携带的参数中不包含模型不同层的时延需求信息，则模型不同层的时延需求信息由 AF 提供。

可选地，处理器 1410，还用于：

确定 AI/ML 模型切分的切分结果之后，向所述待参与模型联合推理的 UE(s)发送第七请求结果，所述第七请求结果中包括 AI/ML 模型切分点信息；

其中，若存在其他参与联合模型推理的 UE(s)，则所述第七请求结果中还

包括其他参与联合模型推理的 UE(s)对应的模型切分点信息,并通过 AF 透传给所述 UE。

可选地,处理器 1410,用于在所述网络实体是应用功能 AF 时,且接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息时,具体包括:

- 5 接收待参与模型联合推理的 UE(s)发送的第八请求,所述第八请求用于请求与 AF 执行模型联合推理操作;所述第八请求中携带的参数包括下述至少一项:与模型标识或模型切分标识关联的分析类型标识、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量;若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,所述第八请求中携带的参数还包括:推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

根据所述第八请求中携带的参数,确定与所述待参与模型联合推理的 UE(s)是否具备完全支持执行模型联合推理的能力;

- 15 若与所述待参与模型联合推理的 UE(s)具备完全支持执行模型联合推理的能力,则将所述第八请求中携带的参数作为所述第一消息;

- 若与所述待参与模型联合推理的 UE(s)不具备完全支持执行模型联合推理的能力,则向网络数据分析功能 NWDAF 发送第九请求,所述第九请求中携带的参数包括所述第八请求中携带的参数,且所述第九请求用于请求
20 NWDAF 通过 NF 查找其他可参与模型推理的 UE(s)以及其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量,若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,所述第九请求中携带的参数还包括:其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

- 25 接收 NWDAF 发送的第九请求结果,并将所述第九请求结果以及所述第八请求中携带的参数作为所述第一消息;其中,所述第九请求结果包括其他可参与模型推理的 UE(s)以及其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量,若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,所述第九请求结果还包括:其他可参与模型推理的 UE(s)
30 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

若第八请求中携带的参数中不包含模型不同层的时延需求信息，则模型不同层的时延需求信息由 AF 提供。

可选地，处理器 1410，还用于：

确定 AI/ML 模型切分的切分结果之后，将确定与所述待参与模型联合推理的 UE(s) 是否具备完全支持执行模型联合推理的能力的结果、所述待参与模型联合推理的 UE(s) 的相关信息以及所述待参与模型联合推理的 UE(s) 对应的模型切分点信息发送给所述 UE；

其中，所述相关信息包括下述至少一项：待参与模型联合推理的 UE(s) 可提供的算力、待参与模型联合推理的 UE(s) 可提供的内存、待参与模型联合推理的 UE(s) 剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述相关信息中还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息。

可选地，处理器 1410，用于在所述网络实体是应用功能 AF 时，且接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息时，具体包括：

向网络数据分析功能 NWDAF 发送第十请求，所述第十请求用于请求 NWDAF 向 5GC NF(s) 采集 UE(s) 对应的用于分析 AI/ML 模型切分的数据，所述第十请求中携带的参数包括下述至少一项：与模型标识或模型切分标识关联的分析类型标识、接收 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足分析条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE(s) 可提供的算力、待参与模型联合推理的 UE(s) 可提供的内存、待参与模型联合推理的 UE(s) 剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第十请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

接收 NWDAF 发送的第十请求结果，所述第十请求结果包括下述至少一项：待参与 AI/ML 模型切分的 UE(s) 或者 SUPI、使用 AI/ML 模型的应用的标识、待参与模型联合推理的 UE(s) 可提供的算力、待参与模型联合推理的 UE(s) 可提供的内存、待参与模型联合推理的 UE(s) 剩余的电量、AI/ML 模型的大小；若第十请求中携带的参数包括推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息，则所述第十请求结果中还包括：推理

AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

将所述第十请求结果作为所述第一消息。

可选地, 处理器 1410, 还用于:

确定 AI/ML 模型切分的切分结果之后, 向待参与模型联合推理的 UE(s)

- 5 发送第十一请求, 所述第十一请求用于请求与所述待参与模型联合推理的 UE(s)执行模型联合推理操作; 其中, 所述第十一请求中携带的参数包括下述至少一项: 模型切分点信息、所述第一消息;

- 接收待参与模型联合推理的 UE(s)发送的第十一请求结果, 所述第十一请求结果是由待参与模型联合推理的 UE(s)根据第十一请求中携带的参数确定的, 所述第十一请求结果包括接受第十一请求或不接受第十一请求。

10 可选地, 处理器 1410, 用于在所述网络实体是新的网络实体 MMF 时, 且接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息时, 具体包括:

- 直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第十二请求, 所述第十二请求用于请求 AI/ML 模型切分分析; 所述第十二请求中携带的参数包括下述至少一项: 与模型标识或模型切分标识关联的分析类型标识、接收 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足分析条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量; 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 所述第十二请求中携带的参数还包括: 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

- 根据所述第十二请求中携带的参数, 向 NWDAF 发送第十三请求, 所述第十三请求中携带的参数包括所述第十二请求中携带的参数, 且所述第十三请求用于请求 NWDAF 向 5GC NF(s)采集 UE(s)的用于分析 AI/ML 模型切分的数据;

- 接收 NWDAF 发送的第十三请求结果; 其中, 所述第十三请求结果包括下述至少一项: 待参与 AI/ML 模型切分的 UE(s)或者 SUPI、使用 AI/ML 模型的应用的标识、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量、

AI/ML 模型的大小；若第十三请求中携带的参数包括推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息，则所述第十二请求结果中还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

5 将所述第十三请求结果作为所述第一消息。

可选地，处理器 1410，还用于：

确定 AI/ML 模型切分的切分结果之后，向所述 AF 发送第八消息，所述第八消息中包括所述 AI/ML 模型切分点、所述第一消息中的数据；所述第八消息用于为 AF 提供向待参与模型联合推理的 UE(s)发送第十四请求时携带的参数，所述第十四请求用于请求与待参与模型联合推理的 UE(s)执行模型联合推理操作，所述第十四请求中携带的参数包括所述第八消息。

可选地，处理器 1410，用于在所述网络实体是新的网络实体 MMF 时，且接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息时，具体包括：

15 直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第十五请求，所述第十五请求用于请求 AI/ML 模型切分分析，所述第十五请求中携带的参数包括下述至少一项：与模型标识或模型切分标识关联的分析类型标识、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第十五请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

根据所述第十五请求中携带的参数，确定 AF 与所述待参与模型联合推理的 UE(s)是否具备完全支持执行模型联合推理的能力；

25 若与所述待参与模型联合推理的 UE(s)具备完全支持执行模型联合推理的能力，则将所述第十五请求中携带的参数作为所述第一消息；

若与所述待参与模型联合推理的 UE(s)不具备完全支持执行模型联合推理的能力，则向 NWDAF 发送第十六请求，所述第十六请求中携带的参数包括所述第十五请求中携带的参数，且所述第十六请求用于请求 NWDAF 通过 NF 查找其他可参与模型推理的 UE(s)以及其他可参与模型推理的 UE(s)可提

供的算力、可提供的内存、剩余的电量，若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第十六请求中携带的参数还包括：其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

- 5 接收 NWDAF 发送的第十六请求结果，并将所述第十六请求结果以及所述第十五请求中携带的参数作为所述第一消息；其中，所述第十六请求结果包括其他可参与模型推理的 UE(s)以及其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量，若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第十六请求结果还包括：其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息

 可选地，处理器 1410，还用于：

 确定 AI/ML 模型切分的切分结果之后，向所述 AF 发送第十五请求结果，所述第十五请求结果中包括 AI/ML 模型切分点信息；

- 15 其中，若存在其他参与联合模型推理的 UE(s)，则所述第十五请求结果中还包括其他参与联合模型推理的 UE(s)对应的模型切分点信息，并通过 AF 透传给所述待参与模型联合推理的 UE(s)。

- 可选地，处理器 1410，用于在所述网络实体是策略控制功能 PCF 时，且接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息时，具体包
20 括：

- 直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第十七请求，所述第十七请求用于请求 AI/ML 模型切分分析；所述第十七请求中携带的参数包括下述至少一项：与模型标识或模型切分标识关联的分析类型标识、接收 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足分析条件的
25 任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第十七请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

- 30 根据所述第十七请求中携带的参数，向网络数据分析功能 NWDAF 发送

第十八请求，所述第十八请求中携带的参数包括所述第十七请求中携带的参数，且所述第十八请求用于请求 NWDAF 向 5GC NF(s)采集 UE(s)对应的第五数据；所述第五数据中包括下述至少一项：待参与 AI/ML 模型切分的 UE(s)或者 SUPI、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量、AI/ML 模型的大小；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第五数据中还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

接收 NWDAF 发送的第十八请求结果；其中，所述第十八请求结果包括所述第五数据；

将所述第十八请求结果作为所述第一消息。

可选地，处理器 1410，还用于：

确定 AI/ML 模型切分的切分结果之后，向所述 AF 发送第九消息，所述第九消息中包括所述 AI/ML 模型切分点、所述第一消息的数据；所述第九消息用于为 AF 提供向待参与模型联合推理的 UE(s)发送第十九请求时携带的参数，所述第十九请求用于请求与待参与模型联合推理的 UE(s)执行模型联合推理操作，所述第十九请求中携带的参数包括所述第九消息。

可选地，处理器 1410，用于在所述网络实体是策略控制功能 PCF 时，且接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息时，具体包括：

直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第二十请求，所述第二十请求用于请求 AI/ML 模型切分分析，所述第二十请求中携带的参数包括下述至少一项：与模型标识或模型切分标识关联的分析类型标识、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第二十请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

根据所述第二十请求中携带的参数，确定所述 AF 以及待参与模型联合推理的 UE 是否具备完全支持执行模型联合推理的能力；

若所述 AF 以及待参与模型联合推理的 UE(s)具备完全支持执行模型联合推理的能力，则将所述第二十请求中携带的参数作为所述第一消息；

若所述 AF 以及待参与模型联合推理的 UE(s)不具备完全支持执行模型联合推理的能力，则向 NWDAF 发送第二十一请求，所述第二十一请求中携带的参数包括所述第二十请求中携带的参数，且所述第二十一请求用于请求 NWDAF 查找其他可参与模型推理的 UE(s)以及其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量，若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第二十一请求中携带的参数还包括：其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

接收 NWDAF 发送的第二十一请求结果，并将所述第二十一请求结果以及所述第二十请求中携带的参数作为所述第一消息；其中，所述第二十一请求结果包括：其他可参与模型推理的 UE(s)以及其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量，若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第二十一请求结果还包括：其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息

若第二十请求中携带的参数中不包含模型不同层的时延需求信息，则模型不同层的时延需求信息由 AF 提供。

可选地，处理器 1410，还用于：

确定 AI/ML 模型切分的切分结果之后，向所述 AF 发送第二十请求结果，所述第二十请求结果中包括 AI/ML 模型切分点信息；

其中，若存在其他参与联合模型推理的 UE(s)，则所述第二十请求结果中还包括其他参与联合模型推理的 UE(s)对应的模型切分点信息，并通过 AF 透传给所述 UE。

在此需要说明的是，本公开提供的辅助模型切分的装置，能够实现图 2-图 12 所示方法实施例所实现的所有方法步骤，且能够达到相同的技术效果，在此不再对本实施例中与方法实施例相同的部分及有益效果进行具体赘述。

图 15 为本公开另一实施例提供的辅助模型切分的装置的结构示意图，如图 15 所示，本实施例提供的辅助模型切分的装置应用于网络实体，则本实施

例提供的辅助模型切分的装置 1500 包括:

接收单元 1501, 用于接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息;

5 确定单元 1502, 用于根据所述第一消息, 确定 AI/ML 模型切分的切分结果。

可选地, 所述第一消息中包括下述至少一项: 待参与 AI/ML 模型切分的 UE(s)或者用户永久标识 SUPI、使用 AI/ML 模型的应用的标识、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量、AI/ML 模型的大小;

10 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 所述第一消息中还包括: 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息。

可选地, 确定单元, 具体用于:

15 根据所述第一消息中的待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量, 确定所述待参与模型联合推理的 UE(s)对应的模型切分点信息;

20 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 则根据所述第一消息中的待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量、推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息, 确定所述待参与模型联合推理的 UE(s)对应的模型切分点信息;

将所述模型切分点信息作为所述切分结果。

可选地, 确定单元, 具体用于:

25 将所有确定参与执行模型联合推理的 UE 中的任一 UE 作为目标 UE, 针对每个所述目标 UE 执行下述步骤:

若所述目标 UE 可提供的算力低于第一预设算力阈值或所述目标 UE 可提供的内存低于第一预设内存阈值或所述目标 UE 剩余的电量低于第一预设电量阈值时, 确定所述目标 UE 执行第一预设数目层的推理;

30 若所述目标 UE 可提供的算力高于第一预设算力阈值并低于第二预设算力阈值, 和所述目标 UE 可提供的内存高于第一预设内存阈值并低于第二预

设内存阈值，以及所述目标 UE 剩余的电量高于第一预设电量阈值并低于第二预设电量阈值时，确定所述目标 UE 执行第二预设数目的推理，第一预设数目层小于第二预设数目层；

若所述目标 UE 可提供的算力高于第二预设算力阈值并低于第三预设算力阈值，和所述目标 UE 可提供的内存高于第二预设内存阈值并低于第三预设内存阈值，以及所述目标 UE 剩余的电量高于第二预设电量阈值并低于第三预设电量阈值时，确定所述目标 UE 执行第三预设数目层的推理，第二预设数目层小于第三预设数目层；以此类推，直至确定所述目标 UE 执行第 N+1 预设数目层的推理，第 N 预设数目层小于第 N+1 预设数目层；

若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，且若推理 AI/ML 模型需要的数据集的隐私等级低于第一预设隐私等级或模型不同层的时延需求信息低于第一预设时延，则所述目标 UE 执行所述第一预设数目层的推理；

若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，且若推理 AI/ML 模型需要的数据集的隐私等级高于第一预设隐私等级并低于第二预设隐私等级，和模型不同层的时延需求信息高于第一预设时延并低于第二预设时延时，确定所述目标 UE 执行第二预设数目层的推理；

若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，且若推理 AI/ML 模型需要的数据集的隐私等级高于第二预设隐私等级并低于第三预设隐私等级，和模型不同层的时延需求信息高于第二预设时延并低于第三预设时延时，确定所述目标 UE 执行第三预设数目层的推理；以此类推，直至确定所述目标 UE 执行第 N+1 预设数目层的推理，第 N 预设数目层小于第 N+1 预设数目层；

根据预设数目层确定 AI/ML 模型切分点信息，所述 AI/ML 模型切分点信息用于表示 AI/ML 模型切分比例，所述预设数目层包括第 N 预设数目层，N 大于或等于一。

可选地，接收单元，具体用于：

如果所述网络实体是网络数据分析功能 NWDAF，直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第一请求，所述第一请求用于请求对 AI/ML 模型切分进行分析；其中，所述第一请求中携带的参数包括下述至少

一项：与模型切分关联的分析类型标识、接受 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足切分条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第一请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

根据所述第一请求中携带的参数，向 5GC NF(s)发送第二消息，所述第二消息用于请求 5GC NF(s)采集 UE(s)对应的第一数据，所述第一数据中包括下述至少一项：待参与 AI/ML 模型切分的 UE(s)以及 SUPI、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量、AI/ML 模型的大小；若第一请求中携带的参数包括推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息，则所述第一数据中还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

接收 5GC NF(s)发送的所述第一数据，并将所述第一数据作为所述第一消息；所述第一数据是由待参与模型联合推理的 UE(s)同意所述第二消息的请求后提供的。

可选地，辅助模型切分的装置，还包括：处理单元；处理单元，用于：确定 AI/ML 模型切分的切分结果之后，向所述 AF 发送第三消息，所述第三消息中包括 AI/ML 模型切分点信息以及所述第一数据；所述第三消息用于为 AF 提供向待参与模型联合推理的 UE(s)发送第二请求时携带的参数，所述第二请求用于请求与待参与模型联合推理的 UE(s)执行模型联合推理操作，所述第二请求中携带的参数包括所述第三消息。

可选地，所述接收单元，具体用于：

如果所述网络实体是网络数据分析功能 NWDAF，直接或通过网络能力开放功能 NEF 接收 AF 发送的第三请求；所述第三请求是由 AF 确定与待参与模型联合推理的 UE(s)具备完全支持执行模型联合推理的能力后确定的，其中，是否具备完全支持执行模型联合推理的能力是由 AF 根据接收到的待参与模型联合推理的 UE(s)发送的第四请求中携带的参数确定的，所述第四请求

用于待参与模型联合推理的 UE(s)请求与 AF 进行模型联合推理,所述第四请求中携带的参数包括下述至少一项:与模型标识或模型切分标识关联的分析类型标识、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量;若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,所述第四请求中携带的参数还包括:推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息,所述第三请求中携带的参数包括所述第四请求中携带的参数;将所述第三请求中携带的参数作为所述第一消息;

其中,若第四请求中携带的参数中不包含模型不同层的时延需求信息,则模型不同层的时延需求信息由 AF 提供。

可选地,辅助模型切分的装置,还包括:处理单元;处理单元,用于:

确定 AI/ML 模型切分的切分结果之后,向所述 AF 发送第四消息,所述第四消息中包括所述 AI/ML 模型切分点、所述第一消息中的数据;所述第四消息用于为 AF 提供向待参与模型联合推理的 UE(s)发送的第四请求结果,所述第四请求结果中包括所述 AI/ML 模型切分点信息。

可选地,接收单元,具体用于:

如果所述网络实体是网络数据分析功能 NWDAF,直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第五请求,所述第五请求是由 AF 确定与待参与模型联合推理的 UE(s)不具备完全支持执行模型联合推理的能力后确定的,其中,是否具备完全支持执行模型联合推理的能力是由 AF 根据接收到的待参与模型联合推理的 UE(s)发送的第六请求确定的,所述第五请求用于请求对 AI/ML 模型切分进行分析以及请求查找其他可参与模型推理的 UE(s)并提供其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量;其中,所述第五请求中携带的参数包括下述至少一项:与模型切分关联的分析类型标识、接受 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足切分条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量;若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,所述第五请求中携带的参数还包括:推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的

时延需求信息；所述第六请求用于待参与模型联合推理的 UE(s)请求与 AF 进行模型联合推理，所述第六请求中携带的参数包括所述第五请求中携带的参数；

5 根据所述第五请求中携带的参数，向 5GC NF(s)发送第五消息，所述第五消息用于请求 5GC NF(s)采集 UE(s)对应的第二数据，所述第二数据中包括下述至少一项：待参与 AI/ML 模型切分的 UE(s)或者 SUPI、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量、AI/ML 模型的大小；若第五请求中携带的参数包括推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息，则所述第二数据中还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

接收 5GC NF(s)发送的所述第二数据，并将所述第二数据作为所述第一消息；所述第二数据是由待参与模型联合推理的 UE(s)同意所述第五消息的请求后提供的；

15 其中，若第六请求中携带的参数中不包含模型不同层的时延需求信息，则模型不同层的时延需求信息由 AF 提供。

可选地，辅助模型切分的装置，还包括处理单元；处理单元，用于：

20 确定 AI/ML 模型切分的切分结果之后，向所述 AF 发送第六消息，所述第六消息中至少包括所述 AI/ML 模型切分点信息以及所述第二数据；所述第六消息用于为 AF 提供向待参与模型联合推理的 UE(s)发送的第六请求结果，所述第六请求结果中包括所述待参与模型联合推理的 UE(s)对应的所述 AI/ML 模型切分点信息、所述其他可参与模型推理的 UE(s)以及对应的所述 AI/ML 模型切分点信息。

可选地，接收单元，具体用于：

25 如果所述网络实体是网络数据分析功能 NWDAF，接收待参与模型联合推理的 UE(s)发送的第七请求，所述第七请求用于请求与 NWDAF 执行模型联合推理操作；所述第七请求中携带的参数包括下述至少一项：待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第七请求中携带的参数还包括：

推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；
根据所述第七请求中携带的参数，确定 AF 与所述待参与模型联合推理的 UE(s)是否具备完全支持执行模型联合推理的能力；

5 若与所述待参与模型联合推理的 UE(s)具备完全支持执行模型联合推理的能力，则将所述第七请求中携带的参数作为所述第一消息；

若与所述待参与模型联合推理的 UE(s)不具备完全支持执行模型联合推理的能力，则请求 NF 查找其他可参与模型推理的 UE(s)并提供其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量，且若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，则请求提供其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

接收 NF 发送的查找结果，所述查找结果包括其他可参与模型推理的 UE(s)以及其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量，若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，则查找结果还包括：其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

将所述第七请求中携带的参数以及所述查找结果作为所述第一消息；

其中，若第七请求中携带的参数中不包含模型不同层的时延需求信息，则模型不同层的时延需求信息由 AF 提供。

20 可选地，辅助模型切分的装置，还包括处理单元；处理单元，用于：

确定 AI/ML 模型切分的切分结果之后，向所述待参与模型联合推理的 UE(s)发送第七请求结果，所述第七请求结果中包括 AI/ML 模型切分点信息；

其中，若存在其他参与联合模型推理的 UE(s)，则所述第七请求结果中还包括其他参与联合模型推理的 UE(s)对应的模型切分点信息，并通过 AF 透传给所述 UE。

25 可选地，接收单元，具体用于：

如果所述网络实体是应用功能 AF，接收待参与模型联合推理的 UE(s)发送的第八请求，所述第八请求用于请求与 AF 执行模型联合推理操作；所述第八请求中携带的参数包括下述至少一项：与模型标识或模型切分标识关联的分析类型标识、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型

联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第八请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；根据所述第八请求中携带的参数，确定与所述待参与模型联合推理的 UE(s)是否具备完全支持执行模型联合推理的能力；

若与所述待参与模型联合推理的 UE(s)具备完全支持执行模型联合推理的能力，则将所述第八请求中携带的参数作为所述第一消息；

若与所述待参与模型联合推理的 UE(s)不具备完全支持执行模型联合推理的能力，则向网络数据分析功能 NWDAF 发送第九请求，所述第九请求中携带的参数包括所述第八请求中携带的参数，且所述第九请求用于请求 NWDAF 通过 NF 查找其他可参与模型推理的 UE(s)以及其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量，若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第九请求中携带的参数还包括：其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

接收 NWDAF 发送的第九请求结果，并将所述第九请求结果以及所述第八请求中携带的参数作为所述第一消息；其中，所述第九请求结果包括其他可参与模型推理的 UE(s)以及其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量，若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第九请求结果还包括：其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

若第八请求中携带的参数中不包含模型不同层的时延需求信息，则模型不同层的时延需求信息由 AF 提供。

可选地，辅助模型切分的装置，还包括处理单元；处理单元，用于：

确定 AI/ML 模型切分的切分结果之后，将确定与所述待参与模型联合推理的 UE(s)是否具备完全支持执行模型联合推理的能力的结果、所述待参与模型联合推理的 UE(s)的相关信息以及所述待参与模型联合推理的 UE(s)对应的模型切分点信息发送给所述 UE；

其中，所述相关信息包括下述至少一项：待参与模型联合推理的 UE(s)

可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述相关信息中还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息。

5 可选地，接收单元，具体用于：

如果所述网络实体是应用功能 AF，向网络数据分析功能 NWDAF 发送第十请求，所述第十请求用于请求 NWDAF 向 5GC NF(s)采集 UE(s)对应的用于分析 AI/ML 模型切分的数据，所述第十请求中携带的参数包括下述至少一项：与模型标识或模型切分标识关联的分析类型标识、接收 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足分析条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第十请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

接收 NWDAF 发送的第十请求结果，所述第十请求结果包括下述至少一项：待参与 AI/ML 模型切分的 UE(s)或者 SUPI、使用 AI/ML 模型的应用的标识、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量、AI/ML 模型的大小；若第十请求中携带的参数包括推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息，则所述第十请求结果中还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

将所述第十请求结果作为所述第一消息。

可选地，辅助模型切分的装置，还包括处理单元；处理单元，用于：

25 确定 AI/ML 模型切分的切分结果之后，向待参与模型联合推理的 UE(s)发送第十一请求，所述第十一请求用于请求与所述待参与模型联合推理的 UE(s)执行模型联合推理操作；其中，所述第十一请求中携带的参数包括下述至少一项：模型切分点信息、所述第一消息；

接收待参与模型联合推理的 UE(s)发送的第十一请求结果，所述第十一请求结果是由待参与模型联合推理的 UE(s)根据第十一请求中携带的参数确定

的, 所述第十一请求结果包括接受第十一请求或不接受第十一请求。

可选地, 接收单元, 具体用于:

如果所述网络实体是新的网络实体 MMF, 直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第十二请求, 所述第十二请求用于请求 AI/ML 模型切分分析; 所述第十二请求中携带的参数包括下述至少一项: 与模型标识或模型切分标识关联的分析类型标识、接收 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足分析条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量; 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 所述第十二请求中携带的参数还包括: 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

根据所述第十二请求中携带的参数, 向 NWDAF 发送第十三请求, 所述第十三请求中携带的参数包括所述第十二请求中携带的参数, 且所述第十三请求用于请求 NWDAF 向 5GC NF(s)采集 UE(s)的用于分析 AI/ML 模型切分的数据;

接收 NWDAF 发送的第十三请求结果; 其中, 所述第十三请求结果包括下述至少一项: 待参与 AI/ML 模型切分的 UE(s)或者 SUPI、使用 AI/ML 模型的应用的标识、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量、AI/ML 模型的大小; 若第十三请求中携带的参数包括推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息, 则所述第十二请求结果中还包括: 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

将所述第十三请求结果作为所述第一消息。

可选地, 辅助模型切分的装置, 还包括处理单元; 处理单元, 用于:

确定 AI/ML 模型切分的切分结果之后, 向所述 AF 发送第八消息, 所述第八消息中包括所述 AI/ML 模型切分点、所述第一消息中的数据; 所述第八消息用于为 AF 提供向待参与模型联合推理的 UE(s)发送第十四请求时携带的参数, 所述第十四请求用于请求与待参与模型联合推理的 UE(s)执行模型联合

推理操作，所述第十四请求中携带的参数包括所述第八消息。

可选地，接收单元，具体用于：

如果所述网络实体是新的网络实体 MMF，直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第十五请求，所述第十五请求用于请求 AI/ML 模型切分分析，所述第十五请求中携带的参数包括下述至少一项：与模型标识或模型切分标识关联的分析类型标识、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第十五请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

根据所述第十五请求中携带的参数，确定 AF 与所述待参与模型联合推理的 UE(s)是否具备完全支持执行模型联合推理的能力；

若与所述待参与模型联合推理的 UE(s)具备完全支持执行模型联合推理的能力，则将所述第十五请求中携带的参数作为所述第一消息；

若与所述待参与模型联合推理的 UE(s)不具备完全支持执行模型联合推理的能力，则向 NWDAF 发送第十六请求，所述第十六请求中携带的参数包括所述第十五请求中携带的参数，且所述第十六请求用于请求 NWDAF 通过 NF 查找其他可参与模型推理的 UE(s)以及其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量，若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第十六请求中携带的参数还包括：其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

接收 NWDAF 发送的第十六请求结果，并将所述第十六请求结果以及所述第十五请求中携带的参数作为所述第一消息；其中，所述第十六请求结果包括其他可参与模型推理的 UE(s)以及其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量，若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第十六请求结果还包括：其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息

可选地，辅助模型切分的装置，还包括处理单元；处理单元，用于：

确定 AI/ML 模型切分的切分结果之后,向所述 AF 发送第十五请求结果,所述第十五请求结果中包括 AI/ML 模型切分点信息;

其中,若存在其他参与联合模型推理的 UE(s),则所述第十五请求结果中还包括其他参与联合模型推理的 UE(s)对应的模型切分点信息,并通过 AF 透

5 传给所述待参与模型联合推理的 UE(s)。

可选地,接收单元,具体用于:

如果所述网络实体是策略控制功能 PCF,直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第十七请求,所述第十七请求用于请求 AI/ML 模型切分分析;所述第十七请求中携带的参数包括下述至少一项:与模型标识或模型切分标识关联的分析类型标识、接收 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足分析条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量;若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,

10 所述第十七请求中携带的参数还包括:推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

根据所述第十七请求中携带的参数,向网络数据分析功能 NWDAF 发送第十八请求,所述第十八请求中携带的参数包括所述第十七请求中携带的参数,且所述第十八请求用于请求 NWDAF 向 5GC NF(s)采集 UE(s)对应的第五数据;所述第五数据中包括下述至少一项:待参与 AI/ML 模型切分的 UE(s)或者 SUPI、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量、AI/ML 模型的大小;若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,所述第五数据中还包括:推理 AI/ML 模型需要的数据集的隐私等级和

20 /或模型不同层的时延需求信息;

接收 NWDAF 发送的第十八请求结果;其中,所述第十八请求结果包括所述第五数据;

将所述第十八请求结果作为所述第一消息。

可选地,辅助模型切分的装置,还包括处理单元;处理单元,用于:

确定 AI/ML 模型切分的切分结果之后,向所述 AF 发送第九消息,所述

30

第九消息中包括所述 AI/ML 模型切分点、所述第一消息的数据；所述第九消息用于为 AF 提供向待参与模型联合推理的 UE(s) 发送第十九请求时携带的参数，所述第十九请求用于请求与待参与模型联合推理的 UE(s) 执行模型联合推理操作，所述第十九请求中携带的参数包括所述第九消息。

5 可选地，接收单元，具体用于：如果所述网络实体是策略控制功能 PCF，

直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第二十请求，所述第二十请求用于请求 AI/ML 模型切分分析，所述第二十请求中携带的参数包括下述至少一项：与模型标识或模型切分标识关联的分析类型标识、待参与模型联合推理的 UE(s) 可提供的算力、待参与模型联合推理的 UE(s) 可提供的内存、待参与模型联合推理的 UE(s) 剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第二十请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

根据所述第二十请求中携带的参数，确定所述 AF 以及待参与模型联合推理的 UE 是否具备完全支持执行模型联合推理的能力；

若所述 AF 以及待参与模型联合推理的 UE(s) 具备完全支持执行模型联合推理的能力，则将所述第二十请求中携带的参数作为所述第一消息；

若所述 AF 以及待参与模型联合推理的 UE(s) 不具备完全支持执行模型联合推理的能力，则向 NWDAF 发送第二十一请求，所述第二十一请求中携带的参数包括所述第二十请求中携带的参数，且所述第二十一请求用于请求 NWDAF 查找其他可参与模型推理的 UE(s) 以及其他可参与模型推理的 UE(s) 可提供的算力、可提供的内存、剩余的电量，若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第二十一请求中携带的参数还包括：其他可参与模型推理的 UE(s) 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

接收 NWDAF 发送的第二十一请求结果，并将所述第二十一请求结果以及所述第二十请求中携带的参数作为所述第一消息；其中，所述第二十一请求结果包括：其他可参与模型推理的 UE(s) 以及其他可参与模型推理的 UE(s) 可提供的算力、可提供的内存、剩余的电量，若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第二十一请求结果还包括：其他

可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息

若第二十请求中携带的参数中不包含模型不同层的时延需求信息，则模型不同层的时延需求信息由 AF 提供。

5 可选地，辅助模型切分的装置，还包括处理单元；处理单元，用于：

确定 AI/ML 模型切分的切分结果之后，向所述 AF 发送第二十请求结果，所述第二十请求结果中包括 AI/ML 模型切分点信息；

其中，若存在其他参与联合模型推理的 UE(s)，则所述第二十请求结果中还包括其他参与联合模型推理的 UE(s)对应的模型切分点信息，并通过 AF 透
10 传给所述 UE。

在此需要说明的是，本公开提供的辅助模型切分的装置，能够实现图 2-图 12 方法实施例所实现的所有方法步骤，且能够达到相同的技术效果，在此不再对本实施例中与方法实施例相同的部分及有益效果进行具体赘述。

图 16 为本公开再一实施例提供的辅助模型切分的装置的结构示意图，如
15 图 16 所示，本实施例提供的辅助模型切分的装置应用于用户设备 UE。则本实施例提供的辅助模型切分的装置包括：收发机 1600，用于在处理器 1610 的控制下接收和发送数据。

其中，在图 16 中，总线架构可以包括任意数量的互联的总线和桥，具体由处理器 1610 代表的一个或多个处理器和存储器 1620 代表的存储器的各种
20 电路链接在一起。总线架构还可以将诸如外围设备、稳压器和功率管理电路等之类的各种其他电路链接在一起，这些都是本领域所公知的，因此，本文不再对其进行进一步描述。总线接口提供接口。收发机 1600 可以是多个元件，即包括发送机和接收机，提供用于在传输介质上与各种其他装置通信的单元，这些传输介质包括无线信道、有线信道、光缆等传输介质。处理器 1610 负责
25 管理总线架构和通常的处理，存储器 1620 可以存储处理器 1610 在执行操作时所使用的数据。

处理器 1610 可以是中央处理器（CPU）、专用集成电路（Application Specific Integrated Circuit, ASIC）、现场可编程门阵列（Field-Programmable Gate Array, FPGA）或复杂可编程逻辑器件（Comple8 Programmable Logic
30 Device, CPLD），处理器也可以采用多核架构。

本实施例中，存储器 1620，用于存储计算机程序；收发机 1600，用于在
处理器的控制下收发数据；处理器 1610，用于读取存储器中的计算机程序并
执行以下操作：

根据自身能力信息，确定 AI/ML 模型切分点信息；

5 将所述 AI/ML 模型切分点信息作为 AI/ML 模型切分的切分结果。

可选地，处理器 1610，用于根据自身能力信息，确定 AI/ML 模型切分点
信息时，具体包括：

根据自身能力信息中的可提供的算力、可提供的内存以及剩余的电量，
确定模型切分点信息；

10 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，则
根据所述自身能力信息中的可提供的算力、可提供的内存、剩余的电量、推
理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息，确
定所述模型切分点信息。

可选地，处理器 1610，用于确定模型切分点信息时，具体包括：

15 将所有确定参与执行模型联合推理的 UE 中的任一 UE 作为目标 UE，针
对每个所述目标 UE 执行下述步骤：

若所述目标 UE 可提供的算力低于第一预设算力阈值或所述目标 UE 可提
供的内存低于第一预设内存阈值或所述目标 UE 剩余的电量低于第一预设电
量阈值时，确定所述目标 UE 执行第一预设数目层的推理；

20 若所述目标 UE 可提供的算力高于第一预设算力阈值并低于第二预设算
力阈值，和所述目标 UE 可提供的内存高于第一预设内存阈值并低于第二预
设内存阈值，以及所述目标 UE 剩余的电量高于第一预设电量阈值并低于第
二预设电量阈值时，确定所述目标 UE 执行第二预设数目的推理，第一
预设数目层小于第二预设数目层；

25 若所述目标 UE 可提供的算力高于第二预设算力阈值并低于第三预设算
力阈值，和所述目标 UE 可提供的内存高于第二预设内存阈值并低于第三预
设内存阈值，以及所述目标 UE 剩余的电量高于第二预设电量阈值并低于第
三预设电量阈值时，确定所述目标 UE 执行第三预设数目层的推理，第二预
设数目层小于第三预设数目层；以此类推，直至确定所述目标 UE 执行第 N+1
30 预设数目层的推理，第 N 预设数目层小于第 N+1 预设数目层；

若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,且若推理 AI/ML 模型需要的数据集的隐私等级低于第一预设隐私等级或模型不同层的时延需求信息低于第一预设时延,则所述目标 UE 执行所述第一预设数目层的推理;

- 5 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,且若推理 AI/ML 模型需要的数据集的隐私等级高于第一预设隐私等级并低于第二预设隐私等级,和模型不同层的时延需求信息高于第一预设时延并低于第二预设时延时,确定所述目标 UE 执行第二预设数目层的推理;

- 10 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,且若推理 AI/ML 模型需要的数据集的隐私等级高于第二预设隐私等级并低于第三预设隐私等级,和模型不同层的时延需求信息高于第二预设时延并低于第三预设时延时,确定所述目标 UE 执行第三预设数目层的推理;以此类推,直至确定所述目标 UE 执行第 N+1 预设数目层的推理,第 N 预设数目层小于第 N+1 预设数目层;

- 15 根据预设数目层确定 AI/ML 模型切分点信息,所述 AI/ML 模型切分点信息用于表示 AI/ML 模型切分比例,所述预设数目层包括第 N 预设数目层, N 大于或等于一。

可选地,处理器 1610,还用于:

- 20 确定 AI/ML 模型切分的切分结果之后,向应用功能 AF 发送第一请求,所述第一请求用于请求与所述 AF 执行模型联合推理操作;其中,所述第一请求中携带的参数包括下述至少一项 AI/ML 模型切分点、与模型标识或模型切分标识关联的分析类型标识、接收 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足分析条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE 可提供的算力、待参与模型联合推理的 UE 可提供的内存、待参与模型联合推理的 UE 剩余的电量;若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,所述第一请求中携带的参数还包括:推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

- 30 接收 AF 发送的第一请求结果,所述第一请求结果是由 AF 根据第一请求中携带的参数确定的,所述第一请求结果包括接受第一请求或不接受第一请

求；

其中，若第一请求中携带的参数中不包含模型不同层的时延需求信息，则模型不同层的时延需求信息由 AF 提供。

可选地，处理器 1610，还用于：

5 确定 AI/ML 模型切分的切分结果之后，通过 AF 向网络数据分析功能 NWDAF 发送第二请求，所述第二请求用于请求与所述 NWDAF 执行模型联合推理操作；其中，所述第二请求中携带的参数包括下述至少一项 AI/ML 模型切分点、与模型标识或模型切分标识关联的分析类型标识、接收 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足分析条件的任意 UEs、
10 AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE 可提供的算力、待参与模型联合推理的 UE 可提供的内存、待参与模型联合推理的 UE 剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第二请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

15 接收 NWDAF 发送的第二请求结果，所述第二请求结果是由 NWDAF 根据第一请求中携带的参数确定的，所述第二请求结果包括接受第一请求或不接受第一请求；

其中，若第二请求中携带的参数中不包含模型不同层的时延需求信息，则模型不同层的时延需求信息由 AF 提供。

20 在此需要说明的是，本公开提供的辅助模型切分的装置，能够实现图 13 所示方法实施例所实现的所有方法步骤，且能够达到相同的技术效果，在此不再对本实施例中与方法实施例相同的部分及有益效果进行具体赘述。

图 17 为本公开又一实施例提供的辅助模型切分的装置的结构示意图，如图 17 所示，本公开实施例提供的辅助模型切分的装置应用于用户设备 UE，
25 则本实施例提供的辅助模型切分的装置 1700 包括：

确定单元 1701，用于根据自身能力信息，确定 AI/ML 模型切分点信息；
处理单元 1702，用于将所述 AI/ML 模型切分点信息作为 AI/ML 模型切分的切分结果。

可选地，确定单元 1701，具体用于：

30 根据自身能力信息中的可提供的算力、可提供的内存以及剩余的电量，

确定模型切分点信息;

若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,则根据所述自身能力信息中的可提供的算力、可提供的内存、剩余的电量、推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息,确定模型切分点信息。

可选地,确定单元 1701,具体用于:

将所有确定参与执行模型联合推理的 UE 中的任一 UE 作为目标 UE,针对每个所述目标 UE 执行下述步骤:

若所述目标 UE 可提供的算力低于第一预设算力阈值或所述目标 UE 可提供的内存低于第一预设内存阈值或所述目标 UE 剩余的电量低于第一预设电量阈值时,确定所述目标 UE 执行第一预设数目层的推理;

若所述目标 UE 可提供的算力高于第一预设算力阈值并低于第二预设算力阈值,和所述目标 UE 可提供的内存高于第一预设内存阈值并低于第二预设内存阈值,以及所述目标 UE 剩余的电量高于第一预设电量阈值并低于第二预设电量阈值时,确定所述目标 UE 执行第二预设数目的推理,第一预设数目层小于第二预设数目层;

若所述目标 UE 可提供的算力高于第二预设算力阈值并低于第三预设算力阈值,和所述目标 UE 可提供的内存高于第二预设内存阈值并低于第三预设内存阈值,以及所述目标 UE 剩余的电量高于第二预设电量阈值并低于第三预设电量阈值时,确定所述目标 UE 执行第三预设数目层的推理,第二预设数目层小于第三预设数目层;以此类推,直至确定所述目标 UE 执行第 N+1 预设数目层的推理,第 N 预设数目层小于第 N+1 预设数目层;

若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,且若推理 AI/ML 模型需要的数据集的隐私等级低于第一预设隐私等级或模型不同层的时延需求信息低于第一预设时延,则所述目标 UE 执行所述第一预设数目层的推理;

若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,且若推理 AI/ML 模型需要的数据集的隐私等级高于第一预设隐私等级并低于第二预设隐私等级,和模型不同层的时延需求信息高于第一预设时延并低于第二预设时延时,确定所述目标 UE 执行第二预设数目层的推理;

若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,且若推理 AI/ML 模型需要的数据集的隐私等级高于第二预设隐私等级并低于第三预设隐私等级,和模型不同层的时延需求信息高于第二预设时延并低于第三预设时延时,确定所述目标 UE 执行第三预设数目层的推理;以此类推,直至确定所述目标 UE 执行第 N+1 预设数目层的推理,第 N 预设数目层小于第 N+1 预设数目层;

根据预设数目层确定 AI/ML 模型切分点信息,所述 AI/ML 模型切分点信息用于表示 AI/ML 模型切分比例,所述预设数目层包括第 N 预设数目层, N 大于或等于一。

可选地,处理单元,还用于:

确定 AI/ML 模型切分的切分结果之后,向应用功能 AF 发送第一请求,所述第一请求用于请求与所述 AF 执行模型联合推理操作;其中,所述第一请求中携带的参数包括下述至少一项 AI/ML 模型切分点、与模型标识或模型切分标识关联的分析类型标识、接收 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足分析条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE 可提供的算力、待参与模型联合推理的 UE 可提供的内存、待参与模型联合推理的 UE 剩余的电量;若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,所述第一请求中携带的参数还包括:推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

接收 AF 发送的第一请求结果,所述第一请求结果是由 AF 根据第一请求中携带的参数确定的,所述第一请求结果包括接受第一请求或不接受第一请求;

其中,若第一请求中携带的参数中不包含模型不同层的时延需求信息,则模型不同层的时延需求信息由 AF 提供。

可选地,处理单元,还用于:

确定 AI/ML 模型切分的切分结果之后,通过 AF 向网络数据分析功能 NWDAF 发送第二请求,所述第二请求用于请求与所述 NWDAF 执行模型联合推理操作;其中,所述第二请求中携带的参数包括下述至少一项 AI/ML 模型切分点、与模型标识或模型切分标识关联的分析类型标识、接收 AI/ML

模型切分的一个用户设备 UE 或一组 UEs 的标识或满足分析条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE 可提供的算力、待参与模型联合推理的 UE 可提供的内存、待参与模型联合推理的 UE 剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第二请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

接收 NWDAF 发送的第二请求结果，所述第二请求结果是由 NWDAF 根据第一请求中携带的参数确定的，所述第二请求结果包括接受第一请求或不接受第一请求；

其中，若第二请求中携带的参数中不包含模型不同层的时延需求信息，则模型不同层的时延需求信息由 AF 提供。

在此需要说明的是，本公开提供的辅助模型切分的装置，能够实现图 13 方法实施例所实现的所有方法步骤，且能够达到相同的技术效果，在此不再对本实施例中与方法实施例相同的部分及有益效果进行具体赘述。

需要说明的是，本公开实施例对单元的划分是示意性的，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式。另外，在本公开各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用，可以存储在一个处理器可读取存储介质中。基于这样的理解，本公开的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的全部或部分可以以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中，包括若干指令用以使得一台计算机设备（可以是个人计算机，服务器，或者网络设备等）或处理器（processor）执行本公开各个实施例方法的全部或部分步骤。而前述的存储介质包括：U 盘、移动硬盘、只读存储器（Read-Only Memory，ROM）、随机存取存储器（Random Access Memory，RAM）、磁碟或者光盘等各种可以存储程序代码的介质。

本公开实施例还提供一种处理器可读存储介质。处理器可读存储介质存储有计算机程序，计算机程序用于使处理器执行上述任一种方法实施例。

其中，处理器可读存储介质可以是处理器能够存取的任何可用介质或数据存储设备，包括但不限于磁性存储器（例如软盘、硬盘、磁带、磁光盘（MO）等）、光学存储器（例如 CD、DVD、BD、HVD 等）、以及半导体存储器（例如 ROM、EPROM、EEPROM、非易失性存储器（NAND FLASH）、固态硬盘（SSD））等。

本领域内的技术人员应明白，本公开的实施例可提供为方法、系统、或计算机程序产品。因此，本公开可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本公开可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质（包括但不限于磁盘存储器和光学存储器等）上实施的计算机程序产品的形式。

本公开是参照根据本公开实施例的方法、设备（系统）、和计算机程序产品的流程图和 / 或方框图来描述的。应理解可由计算机可执行指令实现流程图和 / 或方框图中的每一流程和 / 或方框、以及流程图和 / 或方框图中的流程和 / 或方框的结合。可提供这些计算机可执行指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和 / 或方框图一个方框或多个方框中指定的功能的装置。

这些处理器可执行指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的处理器可读存储器中，使得存储在该处理器可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和 / 或方框图一个方框或多个方框中指定的功能。

这些处理器可执行指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和 / 或方框图一个方框或多个方框中指定的功能的步骤。

显然，本领域的技术人员可以对本公开进行各种改动和变型而不脱离本公开的精神和范围。这样，倘若本公开的这些修改和变型属于本公开权利要求及其等同技术的范围之内，则本公开也意图包含这些改动和变型在内。

权 利 要 求 书

1. 一种辅助模型切分的方法，其特征在于，应用于网络实体，所述方法包括：

接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息；

5 根据所述第一消息，确定 AI/ML 模型切分的切分结果。

2. 根据权利要求 1 所述的方法，其特征在于，所述第一消息中包括下述至少一项：待参与 AI/ML 模型切分的 UE(s)或者用户永久标识 SUPI、使用 AI/ML 模型的应用的标识、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量、AI/ML 模型的大小；

若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第一消息中还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息。

3. 根据权利要求 2 所述的方法，其特征在于，所述根据所述第一消息，确定 AI/ML 模型切分的切分结果，包括：

根据所述第一消息中的待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量，确定所述待参与模型联合推理的 UE(s)对应的模型切分点信息；

若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，则根据所述第一消息中的待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量、推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息，确定所述待参与模型联合推理的 UE(s)对应的模型切分点信息；

将所述模型切分点信息作为所述切分结果。

4. 根据权利要求 3 所述的方法，其特征在于，所述确定所述待参与模型联合推理的 UE(s)对应的模型切分点信息，包括：

将所有确定参与执行模型联合推理的 UE(s)中的任一 UE 作为目标 UE，针对每个所述目标 UE 执行下述步骤：

若所述目标 UE 可提供的算力低于第一预设算力阈值或所述目标 UE 可提供的内存低于第一预设内存阈值或所述目标 UE 剩余的电量低于第一预设电

量阈值时，确定所述目标 UE 执行第一预设数目层的推理；

若所述目标 UE 可提供的算力高于第一预设算力阈值并低于第二预设算力阈值，和所述目标 UE 可提供的内存高于第一预设内存阈值并低于第二预设内存阈值，以及所述目标 UE 剩余的电量高于第一预设电量阈值并低于第二预设电量阈值时，确定所述目标 UE 执行第二预设数目的推理，第一预设数目层小于第二预设数目层；

若所述目标 UE 可提供的算力高于第二预设算力阈值并低于第三预设算力阈值，和所述目标 UE 可提供的内存高于第二预设内存阈值并低于第三预设内存阈值，以及所述目标 UE 剩余的电量高于第二预设电量阈值并低于第三预设电量阈值时，确定所述目标 UE 执行第三预设数目层的推理，第二预设数目层小于第三预设数目层；以此类推，直至确定所述目标 UE 执行第 N+1 预设数目层的推理，第 N 预设数目层小于第 N+1 预设数目层；

若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，且若推理 AI/ML 模型需要的数据集的隐私等级低于第一预设隐私等级或模型不同层的时延需求信息低于第一预设时延，则所述目标 UE 执行所述第一预设数目层的推理；

若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，且若推理 AI/ML 模型需要的数据集的隐私等级高于第一预设隐私等级并低于第二预设隐私等级，和模型不同层的时延需求信息高于第一预设时延并低于第二预设时延时，确定所述目标 UE 执行第二预设数目层的推理；

若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，且若推理 AI/ML 模型需要的数据集的隐私等级高于第二预设隐私等级并低于第三预设隐私等级，和模型不同层的时延需求信息高于第二预设时延并低于第三预设时延时，确定所述目标 UE 执行第三预设数目层的推理；以此类推，直至确定所述目标 UE 执行第 N+1 预设数目层的推理，第 N 预设数目层小于第 N+1 预设数目层；

根据预设数目层确定 AI/ML 模型切分点信息，所述 AI/ML 模型切分点信息用于表示 AI/ML 模型切分比例，所述预设数目层包括第 N 预设数目层，N 大于或等于一。

5. 根据权利要求 1-4 任一项所述的方法，其特征在于，如果所述网络实

体是网络数据分析功能 NWDAF，所述接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息，包括：

直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第一请求，所述第一请求用于请求对 AI/ML 模型切分进行分析；其中，所述第一请求中携带的参数包括下述至少一项：与模型标识或模型切分标识关联的分析类型标识、接受 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足切分条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量；若推理 AI/ML 模型需要数据集的隐私等级和/或模型不同层的时延，所述第一请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

根据所述第一请求中携带的参数，向 5GC NF(s)发送第二消息，所述第二消息用于请求 5GC NF(s)采集 UE(s)对应的第一数据，所述第一数据中包括下述至少一项：待参与 AI/ML 模型切分的 UE(s)以及 SUPI、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量、AI/ML 模型的大小；若第一请求中携带的参数包括推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息，则所述第一数据中还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

接收 5GC NF(s)发送的所述第一数据，并将所述第一数据作为所述第一消息；所述第一数据是由待参与模型联合推理的 UE(s)同意所述第二消息的请求后提供的。

6.根据权利要求 5 所述的方法，其特征在于，所述确定 AI/ML 模型切分的切分结果之后，所述方法还包括：

向所述 AF 发送第三消息，所述第三消息中包括 AI/ML 模型切分点信息以及所述第一数据；所述第三消息用于为 AF 提供向待参与模型联合推理的 UE(s)发送第二请求时携带的参数，所述第二请求用于请求与待参与模型联合推理的 UE(s)执行模型联合推理操作，所述第二请求中携带的参数包括所述第三消息。

7. 根据权利要求 1-4 任一项所述的方法，其特征在于，如果所述网络实

体是网络数据分析功能 NWDAF，所述接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息，包括：

直接或通过网络能力开放功能 NEF 接收 AF 发送的第三请求；所述第三请求是由 AF 确定与待参与模型联合推理的 UE(s)具备完全支持执行模型联合推理的能力后确定的，其中，是否具备完全支持执行模型联合推理的能力是由 AF 根据接收到的待参与模型联合推理的 UE(s)发送的第四请求中携带的参数确定的，所述第四请求用于待参与模型联合推理的 UE(s)请求与 AF 进行模型联合推理，所述第四请求中携带的参数包括下述至少一项：与模型标识或模型切分标识关联的分析类型标识、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第四请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息，所述第三请求中携带的参数包括所述第四请求中携带的参数；

将所述第三请求中携带的参数作为所述第一消息；

其中，若第四请求中携带的参数中不包含模型不同层的时延需求信息，则模型不同层的时延需求信息由 AF 提供。

8. 根据权利要求 7 所述的方法，其特征在于，所述确定 AI/ML 模型切分的切分结果之后，所述方法还包括：

向所述 AF 发送第四消息，所述第四消息中包括所述 AI/ML 模型切分点信息、所述第一消息中的数据；所述第四消息用于为 AF 提供向待参与模型联合推理的 UE(s)发送的第四请求结果，所述第四请求结果中包括所述 AI/ML 模型切分点信息。

9. 根据权利要求 1-4 任一项所述的方法，其特征在于，如果所述网络实体是网络数据分析功能 NWDAF，所述接收用于请求人工智能/机器学习 AI/ML 模型切分分析的第一消息，包括：

直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第五请求，所述第五请求是由 AF 确定与待参与模型联合推理的 UE(s)不具备完全支持执行模型联合推理的能力后确定的，其中，是否具备完全支持执行模型联合推理的能力是由 AF 根据接收到的待参与模型联合推理的 UE(s)发送的第六请求

确定的, 所述第五请求用于请求对 AI/ML 模型切分进行分析以及请求查找其他可参与模型推理的 UE(s)并提供其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量; 其中, 所述第五请求中携带的参数包括下述至少一项: 与模型切分关联的分析类型标识、接受 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足切分条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量; 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 所述第五请求中携带的参数还包括: 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息; 所述第六请求用于待参与模型联合推理的 UE(s)请求与 AF 进行模型联合推理, 所述第六请求中携带的参数包括所述第五请求中携带的参数;

根据所述第五请求中携带的参数, 向 5GC NF(s)发送第五消息, 所述第五消息用于请求 5GC NF(s)采集 UE(s)对应的第二数据, 所述第二数据中包括下述至少一项: 待参与 AI/ML 模型切分的 UE(s)或者 SUPI、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量、AI/ML 模型的大小; 若第五请求中携带的参数包括推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息, 则所述第二数据中还包括: 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

接收 5GC NF(s)发送的所述第二数据, 并将所述第二数据作为所述第一消息; 所述第二数据是由待参与模型联合推理的 UE(s)同意所述第五消息的请求后提供的;

其中, 若第六请求中携带的参数中不包含模型不同层的时延需求信息, 则模型不同层的时延需求信息由 AF 提供。

10. 根据权利要求 9 所述的方法, 其特征在于, 所述确定 AI/ML 模型切分的切分结果之后, 所述方法还包括:

向所述 AF 发送第六消息, 所述第六消息中至少包括所述 AI/ML 模型切分点信息以及所述第二数据; 所述第六消息用于为 AF 提供向待参与模型联合推理的 UE(s)发送的第六请求结果, 所述第六请求结果中包括所述待参与模

型联合推理的 UE(s)以及对应的所述 AI/ML 模型切分点信息。

11. 根据权利要求 1-4 任一项所述的方法,其特征在于,如果所述网络实体是网络数据分析功能 NWDAF,所述接收用于请求人工智能/机器学习 AI/ML 模型切分分析的第一消息,包括:

5 接收待参与模型联合推理的 UE(s)发送的第七请求,所述第七请求用于请求与 NWDAF 执行模型联合推理操作;所述第七请求中携带的参数包括下述至少一项:待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量;若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,所述第七请求
10 中携带的参数还包括:推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;根据所述第七请求中携带的参数,确定 AF 与所述待参与模型联合推理的 UE(s)是否具备完全支持执行模型联合推理的能力;

若与所述待参与模型联合推理的 UE(s)具备完全支持执行模型联合推理的能力,则将所述第七请求中携带的参数作为所述第一消息;

15 若与所述待参与模型联合推理的 UE(s)不具备完全支持执行模型联合推理的能力,则请求 NF 查找其他可参与模型推理的 UE(s)并提供其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量,且若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,则请求提供其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同
20 层的时延需求信息;

接收 NF 发送的查找结果,所述查找结果包括其他可参与模型推理的 UE(s)以及其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量,若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,则查找结果还包括:其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;
25

将所述第七请求中携带的参数以及所述查找结果作为所述第一消息;

其中,若第七请求中携带的参数中不包含模型不同层的时延需求信息,则模型不同层的时延需求信息由 AF 提供。

12. 根据权利要求 11 所述的方法,其特征在于,所述确定 AI/ML 模型切
30 分的切分结果之后,所述方法还包括:

向所述待参与模型联合推理的 UE(s) 发送第七请求结果, 所述第七请求结果中包括 AI/ML 模型切分点信息;

其中, 若存在其他参与联合模型推理的 UE(s), 则所述第七请求结果中还包括其他参与联合模型推理的 UE(s) 对应的模型切分点信息, 并通过 AF 透传给所述 UE(s)。

13. 根据权利要求 1-4 任一项所述的方法, 其特征在于, 如果所述网络实体是应用功能 AF, 所述接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息, 包括:

接收待参与模型联合推理的 UE(s) 发送的第八请求, 所述第八请求用于请求与 AF 执行模型联合推理操作; 所述第八请求中携带的参数包括下述至少一项: 与模型标识或模型切分标识关联的分析类型标识、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE(s) 可提供的算力、待参与模型联合推理的 UE(s) 可提供的内存、待参与模型联合推理的 UE(s) 剩余的电量; 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 所述第八请求中携带的参数还包括: 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息; 根据所述第八请求中携带的参数, 确定与所述待参与模型联合推理的 UE(s) 是否具备完全支持执行模型联合推理的能力;

若与所述待参与模型联合推理的 UE(s) 具备完全支持执行模型联合推理的能力, 则将所述第八请求中携带的参数作为所述第一消息;

若与所述待参与模型联合推理的 UE(s) 不具备完全支持执行模型联合推理的能力, 则向网络数据分析功能 NWDAF 发送第九请求, 所述第九请求中携带的参数包括所述第八请求中携带的参数, 且所述第九请求用于请求 NWDAF 通过 NF 查找其他可参与模型推理的 UE(s) 以及其他可参与模型推理的 UE(s) 可提供的算力、可提供的内存、剩余的电量, 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 所述第九请求中携带的参数还包括: 其他可参与模型推理的 UE(s) 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

接收 NWDAF 发送的第九请求结果, 并将所述第九请求结果以及所述第八请求中携带的参数作为所述第一消息; 其中, 所述第九请求结果包括其他

可参与模型推理的 UE(s)以及其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量,若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,所述第九请求结果还包括:其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

5 若第八请求中携带的参数中不包含模型不同层的时延需求信息,则模型不同层的时延需求信息由 AF 提供。

14. 根据权利要求 13 所述的方法,其特征在于,所述确定 AI/ML 模型切分的切分结果之后,所述方法还包括:

10 将确定与所述待参与模型联合推理的 UE(s)是否具备完全支持执行模型联合推理的能力的结果、所述待参与模型联合推理的 UE(s)的相关信息以及所述待参与模型联合推理的 UE(s)对应的模型切分点信息发送给所述 UE;

其中,所述相关信息包括下述至少一项:待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量;若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,所述相关信息中还包括:推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息。

15 15. 根据权利要求 1-4 任一项所述的方法,其特征在于,如果所述网络实体是应用功能 AF,所述接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息,包括:

20 向网络数据分析功能 NWDAF 发送第十请求,所述第十请求用于请求 NWDAF 向 5GC NF(s)采集 UE(s)对应的用于分析 AI/ML 模型切分的数据,所述第十请求中携带的参数包括下述至少一项:与模型标识或模型切分标识关联的分析类型标识、接收 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足分析条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量;若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,所述第十请求中携带的参数还包括:推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

30 接收 NWDAF 发送的第十请求结果,所述第十请求结果包括下述至少一

项：待参与 AI/ML 模型切分的 UE(s)或者 SUPI、使用 AI/ML 模型的应用的标识、待参与模型联合推理的 UE 可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量、AI/ML 模型的大小；若第十请求中携带的参数包括推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息，则所述第十请求结果中还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

将所述第十请求结果作为所述第一消息。

16. 根据权利要求 15 所述的方法，其特征在于，所述确定 AI/ML 模型切分的切分结果之后，所述方法还包括：

10 向待参与模型联合推理的 UE(s)发送第十一请求，所述第十一请求用于请求与所述待参与模型联合推理的 UE(s)执行模型联合推理操作；其中，所述第十一请求中携带的参数包括下述至少一项：模型切分点信息、所述第一消息；

接收待参与模型联合推理的 UE(s)发送的第十一请求结果，所述第十一请求结果是由待参与模型联合推理的 UE(s)根据第十一请求中携带的参数确定的，所述第十一请求结果包括接受第十一请求或不接受第十一请求。

17. 根据权利要求 1-4 任一项所述的方法，其特征在于，如果所述网络实体是新的网络实体 MMF，所述接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息，包括：

直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第十二请求，所述第十二请求用于请求 AI/ML 模型切分分析；所述第十二请求中携带的参数包括下述至少一项：与模型标识或模型切分标识关联的分析类型标识、接收 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足分析条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第十二请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

根据所述第十二请求中携带的参数，向 NWDAF 发送第十三请求，所述第十三请求中携带的参数包括所述第十二请求中携带的参数，且所述第十三请求用于请求 NWDAF 向 5GC NF(s)采集 UE(s)的用于分析 AI/ML 模型切分

的数据;

接收 NWDAF 发送的第十三请求结果; 其中, 所述第十三请求结果包括下述至少一项: 待参与 AI/ML 模型切分的 UE(s)或者 SUPI、使用 AI/ML 模型的应用的标识、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量、AI/ML 模型的大小; 若第十三请求中携带的参数包括推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息, 则所述第十二请求结果中还包括: 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

10 将所述第十三请求结果作为所述第一消息。

18. 根据权利要求 17 所述的方法, 其特征在于, 所述确定 AI/ML 模型切分的切分结果之后, 所述方法还包括:

向所述 AF 发送第八消息, 所述第八消息中包括所述 AI/ML 模型切分点、所述第一消息中的数据; 所述第八消息用于为 AF 提供向待参与模型联合推理的 UE(s)发送第十四请求时携带的参数, 所述第十四请求用于请求与待参与模型联合推理的 UE(s)执行模型联合推理操作, 所述第十四请求中携带的参数包括所述第八消息。

19. 根据权利要求 1-4 任一项所述的方法, 其特征在于, 如果所述网络实体是新的网络实体 MMF, 所述接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息, 包括:

直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第十五请求, 所述第十五请求用于请求 AI/ML 模型切分分析, 所述第十五请求中携带的参数包括下述至少一项: 与模型标识或模型切分标识关联的分析类型标识、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量; 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 所述第十五请求中携带的参数还包括: 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

根据所述第十五请求中携带的参数, 确定 AF 与所述待参与模型联合推理的 UE(s)是否具备完全支持执行模型联合推理的能力;

若与所述待参与模型联合推理的 UE(s)具备完全支持执行模型联合推理的能力，则将所述第十五请求中携带的参数作为所述第一消息；

若与所述待参与模型联合推理的 UE(s)不具备完全支持执行模型联合推理的能力，则向 NWDAF 发送第十六请求，所述第十六请求中携带的参数包括所述第十五请求中携带的参数，且所述第十六请求用于请求 NWDAF 通过 NF 查找其他可参与模型推理的 UE(s)以及其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量，若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第十六请求中携带的参数还包括：其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

接收 NWDAF 发送的第十六请求结果，并将所述第十六请求结果以及所述第十五请求中携带的参数作为所述第一消息；其中，所述第十六请求结果包括其他可参与模型推理的 UE(s)以及其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量，若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第十六请求结果还包括：其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息。

20. 根据权利要求 19 所述的方法，其特征在于，所述确定 AI/ML 模型切分的切分结果之后，所述方法还包括：

向所述 AF 发送第十五请求结果，所述第十五请求结果中包括 AI/ML 模型切分点信息；

其中，若存在其他参与联合模型推理的 UE(s)，则所述第十五请求结果中还包括其他参与联合模型推理的 UE(s)对应的模型切分点信息，并通过 AF 透传给所述待参与模型联合推理的 UE(s)。

21. 根据权利要求 1-4 任一项所述的方法，其特征在于，如果所述网络实体是策略控制功能 PCF，所述接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息，包括：

直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第十七请求，所述第十七请求用于请求 AI/ML 模型切分分析；所述第十七请求中携带的参数包括下述至少一项：与模型标识或模型切分标识关联的分析类型标识、接

收 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足分析条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第十七请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

根据所述第十七请求中携带的参数，向网络数据分析功能 NWDAF 发送第十八请求，所述第十八请求中携带的参数包括所述第十七请求中携带的参数，且所述第十八请求用于请求 NWDAF 向 5GC NF(s)采集 UE(s)对应的第五数据；所述第五数据中包括下述至少一项：待参与 AI/ML 模型切分的 UE(s)或者 SUPI、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量、AI/ML 模型的大小；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第五数据中还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

接收 NWDAF 发送的第十八请求结果；其中，所述第十八请求结果包括所述第五数据；

将所述第十八请求结果作为所述第一消息。

22. 根据权利要求 21 所述的方法，其特征在于，所述确定 AI/ML 模型切分的切分结果之后，所述方法还包括：

向所述 AF 发送第九消息，所述第九消息中包括所述 AI/ML 模型切分点、所述第一消息的数据；所述第九消息用于为 AF 提供向待参与模型联合推理的 UE(s)发送第十九请求时携带的参数，所述第十九请求用于请求与待参与模型联合推理的 UE(s)执行模型联合推理操作，所述第十九请求中携带的参数包括所述第九消息。

23. 根据权利要求 1-4 任一项所述的方法，其特征在于，如果所述网络实体是策略控制功能 PCF，所述接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息，包括：

直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第二十请求，所述第二十请求用于请求 AI/ML 模型切分分析，所述第二十请求中携带的参

数包括下述至少一项：与模型标识或模型切分标识关联的分析类型标识、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第二十请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

根据所述第二十请求中携带的参数，确定所述 AF 以及待参与模型联合推理的 UE(s)是否具备完全支持执行模型联合推理的能力；

若所述 AF 以及待参与模型联合推理的 UE(s)具备完全支持执行模型联合推理的能力，则将所述第二十请求中携带的参数作为所述第一消息；

若所述 AF 以及待参与模型联合推理的 UE(s)不具备完全支持执行模型联合推理的能力，则向 NWDAF 发送第二十一请求，所述第二十一请求中携带的参数包括所述第二十请求中携带的参数，且所述第二十一请求用于请求 NWDAF 通过 NF 查找其他可参与模型推理的 UE(s)以及其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量，若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第二十一请求中携带的参数还包括：其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

接收 NWDAF 发送的第二十一请求结果，并将所述第二十一请求结果以及所述第二十请求中携带的参数作为所述第一消息；其中，所述第二十一请求结果包括：其他可参与模型推理的 UE(s)以及其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量，若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第二十一请求结果还包括：其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

若第二十请求中携带的参数中不包含模型不同层的时延需求信息，则模型不同层的时延需求信息由 AF 提供。

24. 根据权利要求 23 所述的方法，其特征在于，所述确定 AI/ML 模型切分的切分结果之后，所述方法还包括：

向所述 AF 发送第二十请求结果，所述第二十请求结果中包括 AI/ML 模

型切分点信息;

其中,若存在其他参与联合模型推理的 UE(s),则所述第二十请求结果中还包括其他参与联合模型推理的 UE(s)对应的模型切分点信息,并通过 AF 透传给所述 UE(s)。

- 5 25. 一种辅助模型切分的方法,其特征在于,应用于用户设备 UE,所述方法包括:

根据自身能力信息,确定 AI/ML 模型切分点信息;

将所述 AI/ML 模型切分点信息作为 AI/ML 模型切分的切分结果。

- 10 26. 根据权利要求 25 所述的方法,其特征在于,所述根据自身能力信息,确定 AI/ML 模型切分点信息,包括:

根据自身能力信息中的可提供的算力、可提供的内存以及剩余的电量,确定模型切分点信息;

- 15 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,则根据所述自身能力信息中的可提供的算力、可提供的内存、剩余的电量、推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息,确定模型切分点信息。

27. 根据权利要求 26 所述的方法,其特征在于,所述确定模型切分点信息,包括:

- 20 将所有确定参与执行模型联合推理的 UE(s)中的任一 UE 作为目标 UE,针对每个所述目标 UE 执行下述步骤:

若所述目标 UE 可提供的算力低于第一预设算力阈值或所述目标 UE 可提供的内存低于第一预设内存阈值或所述目标 UE 剩余的电量低于第一预设电量阈值时,确定所述目标 UE 执行第一预设数目层的推理;

- 25 若所述目标 UE 可提供的算力高于第一预设算力阈值并低于第二预设算力阈值,和所述目标 UE 可提供的内存高于第一预设内存阈值并低于第二预设内存阈值,以及所述目标 UE 剩余的电量高于第一预设电量阈值并低于第二预设电量阈值时,确定所述目标 UE 执行第二预设数目的推理,第一预设数目层小于第二预设数目层;

- 30 若所述目标 UE 可提供的算力高于第二预设算力阈值并低于第三预设算力阈值,和所述目标 UE 可提供的内存高于第二预设内存阈值并低于第三预

设内存阈值, 以及所述目标 UE 剩余的电量高于第二预设电量阈值并低于第三预设电量阈值时, 确定所述目标 UE 执行第三预设数目层的推理, 第二预设数目层小于第三预设数目层; 以此类推, 直至确定所述目标 UE 执行第 N+1 预设数目层的推理, 第 N 预设数目层小于第 N+1 预设数目层;

5 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 且若推理 AI/ML 模型需要的数据集的隐私等级低于第一预设隐私等级或模型不同层的时延需求信息低于第一预设时延, 则所述目标 UE 执行所述第一预设数目层的推理;

 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 且
10 若推理 AI/ML 模型需要的数据集的隐私等级高于第一预设隐私等级并低于第二预设隐私等级, 和模型不同层的时延需求信息高于第一预设时延并低于第二预设时延时, 确定所述目标 UE 执行第二预设数目层的推理;

 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 且
15 若推理 AI/ML 模型需要的数据集的隐私等级高于第二预设隐私等级并低于第三预设隐私等级, 和模型不同层的时延需求信息高于第二预设时延并低于第三预设时延时, 确定所述目标 UE 执行第三预设数目层的推理; 以此类推, 直至确定所述目标 UE 执行第 N+1 预设数目层的推理, 第 N 预设数目层小于第 N+1 预设数目层;

 根据预设数目层确定 AI/ML 模型切分点信息, 所述 AI/ML 模型切分点信息用于表示 AI/ML 模型切分比例, 所述预设数目层包括第 N 预设数目层, N 大于或等于一。

28. 根据权利要求 27 所述的方法, 其特征在于, 所述确定 AI/ML 模型切分的切分结果之后, 所述方法还包括:

 向应用功能 AF 发送第一请求, 所述第一请求用于请求与所述 AF 执行模
25 型联合推理操作; 其中, 所述第一请求中携带的参数包括下述至少一项 AI/ML 模型切分点、与模型标识或模型切分标识关联的分析类型标识、接收 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足分析条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE 可提供的算力、待参与模型联合推理的 UE 可提供的内存、待参与模型联合推
30 理的 UE 剩余的电量; 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型

不同层的时延，所述第一请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

接收 AF 发送的第一请求结果，所述第一请求结果是由 AF 根据第一请求中携带的参数确定的，所述第一请求结果包括接受第一请求或不接受第一请求；

其中，若第一请求中携带的参数中不包含模型不同层的时延需求信息，则模型不同层的时延需求信息由 AF 提供。

29. 根据权利要求 27 所述的方法，其特征在于，所述确定 AI/ML 模型切分的切分结果之后，所述方法还包括：

通过 AF 向网络数据分析功能 NWDAF 发送第二请求，所述第二请求用于请求与所述 NWDAF 执行模型联合推理操作；其中，所述第二请求中携带的参数包括下述至少一项 AI/ML 模型切分点、与模型标识或模型切分标识关联的分析类型标识、接收 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足分析条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE 可提供的算力、待参与模型联合推理的 UE 可提供的内存、待参与模型联合推理的 UE 剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第二请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

接收 NWDAF 发送的第二请求结果，所述第二请求结果是由 NWDAF 根据第一请求中携带的参数确定的，所述第二请求结果包括接受第一请求或不接受第一请求；

其中，若第二请求中携带的参数中不包含模型不同层的时延需求信息，则模型不同层的时延需求信息由 AF 提供。

30. 一种辅助模型切分的装置，其特征在于，应用于网络实体，所述装置包括存储器，收发机，处理器：

存储器，用于存储计算机程序；收发机，用于在所述处理器的控制下收发数据；处理器，用于读取所述存储器中的计算机程序并执行以下操作：

接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息；

根据所述第一消息，确定 AI/ML 模型切分的切分结果。

31. 一种辅助模型切分的装置，其特征在于，应用于用户设备 UE，所述装置包括存储器，收发机，处理器：

存储器，用于存储计算机程序；收发机，用于在所述处理器的控制下收发数据；处理器，用于读取所述存储器中的计算机程序并执行以下操作：

5 根据自身能力信息，确定 AI/ML 模型切分点信息；

将所述 AI/ML 模型切分点信息作为 AI/ML 模型切分的切分结果。

32. 一种辅助模型切分的装置，其特征在于，应用于网络实体，所述装置包括：

10 接收单元，用于接收用于辅助人工智能/机器学习 AI/ML 模型切分分析的第一消息；

确定单元，用于根据所述第一消息，确定 AI/ML 模型切分的切分结果。

33. 根据权利要求 32 所述的装置，其特征在于，所述第一消息中包括下述至少一项：待参与 AI/ML 模型切分的 UE(s)或者用户永久标识 SUPI、使用 AI/ML 模型的应用的标识、待参与模型联合推理的 UE(s)可提供的算力、待
15 参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量、AI/ML 模型的大小；

若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第一消息中还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息。

20 34. 根据权利要求 32 或 33 所述的装置，其特征在于，接收单元，具体用于：

如果所述网络实体是网络数据分析功能 NWDAF，直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第一请求，所述第一请求用于请求对 AI/ML 模型切分进行分析；其中，所述第一请求中携带的参数包括下述至少
25 一项：与模型标识或模型切分标识关联的分析类型标识、接受 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足切分条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量；若推理 AI/ML 模型需要数据集的隐私等级和/或模型不同
30 层的时延，所述第一请求中携带的参数还包括：推理 AI/ML 模型需要的数据

集的隐私等级和/或模型不同层的时延需求信息;

根据所述第一请求中携带的参数,向 5GC NF(s)发送第二消息,所述第二消息用于请求 5GC NF(s)采集 UE(s)对应的第一数据,所述第一数据中包括下述至少一项:待参与 AI/ML 模型切分的 UE(s)以及 SUPI、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量、AI/ML 模型的大小;若第一请求中携带的参数包括推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息,则所述第一数据中还包括:推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

接收 5GC NF(s)发送的所述第一数据,并将所述第一数据作为所述第一消息;所述第一数据是由待参与模型联合推理的 UE(s)同意所述第二消息的请求后提供的。

35. 根据权利要求 32 或 33 所述的装置,其特征在于,接收单元,具体用于:

如果所述网络实体是网络数据分析功能 NWDAF,直接或通过网络能力开放功能 NEF 接收 AF 发送的第三请求;所述第三请求是由 AF 确定与待参与模型联合推理的 UE(s)具备完全支持执行模型联合推理的能力后确定的,其中,是否具备完全支持执行模型联合推理的能力是由 AF 根据接收到的待参与模型联合推理的 UE(s)发送的第四请求中携带的参数确定的,所述第四请求用于待参与模型联合推理的 UE(s)请求与 AF 进行模型联合推理,所述第四请求中携带的参数包括下述至少一项:与模型标识或模型切分标识关联的分析类型标识、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量;若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,所述第四请求中携带的参数还包括:推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息,所述第三请求中携带的参数包括所述第四请求中携带的参数;

将所述第三请求中携带的参数作为所述第一消息;

其中,若第四请求中携带的参数中不包含模型不同层的时延需求信息,则模型不同层的时延需求信息由 AF 提供。

36. 根据权利要求 32 或 33 所述的装置，其特征在于，发送单元，具体用于：

如果所述网络实体是网络数据分析功能 NWDAF，直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第五请求，所述第五请求是由 AF 确定与待参与模型联合推理的 UE(s) 不具备完全支持执行模型联合推理的能力后确定的，其中，是否具备完全支持执行模型联合推理的能力是由 AF 根据接收到的待参与模型联合推理的 UE(s) 发送的第六请求确定的，所述第五请求用于请求对 AI/ML 模型切分进行分析以及请求查找其他可参与模型推理的 UE(s) 并提供其他可参与模型推理的 UE(s) 可提供的算力、可提供的内存、剩余的电量；其中，所述第五请求中携带的参数包括下述至少一项：与模型切分关联的分析类型标识、接受 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足切分条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE(s) 可提供的算力、待参与模型联合推理的 UE(s) 可提供的内存、待参与模型联合推理的 UE(s) 剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第五请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；所述第六请求用于待参与模型联合推理的 UE(s) 请求与 AF 进行模型联合推理，所述第六请求中携带的参数包括所述第五请求中携带的参数；

根据所述第五请求中携带的参数，向 5GC NF(s) 发送第五消息，所述第五消息用于请求 5GC NF(s) 采集 UE(s) 对应的第二数据，所述第二数据中包括下述至少一项：待参与 AI/ML 模型切分的 UE(s) 或者 SUPI、待参与模型联合推理的 UE(s) 可提供的算力、待参与模型联合推理的 UE(s) 可提供的内存、待参与模型联合推理的 UE(s) 剩余的电量、AI/ML 模型的大小；若第五请求中携带的参数包括推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息，则所述第二数据中还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

接收 5GC NF(s) 发送的所述第二数据，并将所述第二数据作为所述第一消息；所述第二数据是由待参与模型联合推理的 UE(s) 同意所述第五消息的请求后提供的；

其中，若第六请求中携带的参数中不包含模型不同层的时延需求信息，则模型不同层的时延需求信息由 AF 提供。

37. 根据权利要求 32 或 33 所述的装置，其特征在于，接收单元，具体用于：

5 如果所述网络实体是网络数据分析功能 NWDAF，接收待参与模型联合推理的 UE(s)发送的第七请求，所述第七请求用于请求与 NWDAF 执行模型联合推理操作；所述第七请求中携带的参数包括下述至少一项：待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量；若推理 AI/ML 模型时需要数据集
10 的隐私等级和/或模型不同层的时延，所述第七请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；根据所述第七请求中携带的参数，确定 AF 与所述待参与模型联合推理的 UE(s)是否具备完全支持执行模型联合推理的能力；

 若与所述待参与模型联合推理的 UE(s)具备完全支持执行模型联合推理
15 的能力，则将所述第七请求中携带的参数作为所述第一消息；

 若与所述待参与模型联合推理的 UE(s)不具备完全支持执行模型联合推理的能力，则请求 NF 查找其他可参与模型推理的 UE(s)并提供其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量，且若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，则请求提供其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同
20 层的时延需求信息；

 接收 NF 发送的查找结果，所述查找结果包括其他可参与模型推理的 UE(s)以及其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量，若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，
25 则查找结果还包括：其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

 将所述第七请求中携带的参数以及所述查找结果作为所述第一消息；

 其中，若第七请求中携带的参数中不包含模型不同层的时延需求信息，则模型不同层的时延需求信息由 AF 提供。

30 38. 根据权利要求 32 或 33 所述的装置，其特征在于，接收单元，具体

用于:

如果所述网络实体是应用功能 AF, 接收待参与模型联合推理的 UE(s) 发送的第八请求, 所述第八请求用于请求与 AF 执行模型联合推理操作; 所述第八请求中携带的参数包括下述至少一项: 与模型标识或模型切分标识关联的分析类型标识、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE(s) 可提供的算力、待参与模型联合推理的 UE(s) 可提供的内存、待参与模型联合推理的 UE(s) 剩余的电量; 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 所述第八请求中携带的参数还包括: 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息; 根据所述第八请求中携带的参数, 确定与所述待参与模型联合推理的 UE(s) 是否具备完全支持执行模型联合推理的能力;

若与所述待参与模型联合推理的 UE(s) 具备完全支持执行模型联合推理的能力, 则将所述第八请求中携带的参数作为所述第一消息;

若与所述待参与模型联合推理的 UE(s) 不具备完全支持执行模型联合推理的能力, 则向网络数据分析功能 NWDAF 发送第九请求, 所述第九请求中携带的参数包括所述第八请求中携带的参数, 且所述第九请求用于请求 NWDAF 通过 NF 查找其他可参与模型推理的 UE(s) 以及其他可参与模型推理的 UE(s) 可提供的算力、可提供的内存、剩余的电量, 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 所述第九请求中携带的参数还包括: 其他可参与模型推理的 UE(s) 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

接收 NWDAF 发送的第九请求结果, 并将所述第九请求结果以及所述第八请求中携带的参数作为所述第一消息; 其中, 所述第九请求结果包括其他可参与模型推理的 UE(s) 以及其他可参与模型推理的 UE(s) 可提供的算力、可提供的内存、剩余的电量, 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 所述第九请求结果还包括: 其他可参与模型推理的 UE(s) 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

若第八请求中携带的参数中不包含模型不同层的时延需求信息, 则模型不同层的时延需求信息由 AF 提供。

39. 根据权利要求 32 或 33 所述的装置, 其特征在于, 接收单元, 具体

用于:

如果所述网络实体是应用功能 AF, 向网络数据分析功能 NWDAF 发送第十请求, 所述第十请求用于请求 NWDAF 向 5GC NF(s) 采集 UE(s) 对应的用于分析 AI/ML 模型切分的数据, 所述第十请求中携带的参数包括下述至少一项:

5 与模型标识或模型切分标识关联的分析类型标识、接收 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足分析条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE(s) 可提供的算力、待参与模型联合推理的 UE(s) 可提供的内存、待参与模型联合推理的 UE(s) 剩余的电量; 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 所述第十请求中携带的参数还包括: 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

接收 NWDAF 发送的第十请求结果, 所述第十请求结果包括下述至少一项: 待参与 AI/ML 模型切分的 UE(s) 或者 SUPI、使用 AI/ML 模型的应用的标识、待参与模型联合推理的 UE 可提供的算力、待参与模型联合推理的 UE(s) 可提供的内存、待参与模型联合推理的 UE(s) 剩余的电量、AI/ML 模型的大小; 若第十请求中携带的参数包括推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息, 则所述第十请求结果中还包括: 推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

将所述第十请求结果作为所述第一消息。

20 40. 根据权利要求 32 或 33 所述的装置, 其特征在于, 接收单元, 具体用于:

如果所述网络实体是新的网络实体 MMF, 直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第十二请求, 所述第十二请求用于请求 AI/ML 模型切分分析; 所述第十二请求中携带的参数包括下述至少一项: 与模型标识或模型切分标识关联的分析类型标识、接收 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足分析条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE(s) 可提供的算力、待参与模型联合推理的 UE(s) 可提供的内存、待参与模型联合推理的 UE(s) 剩余的电量; 若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延, 所述第十二请求中携带的参数还包括: 推理 AI/ML 模型需要的数据集的隐私

25 30

等级和/或模型不同层的时延需求信息;

根据所述第十二请求中携带的参数,向 NWDAF 发送第十三请求,所述第十三请求中携带的参数包括所述第十二请求中携带的参数,且所述第十三请求用于请求 NWDAF 向 5GC NF(s)采集 UE(s)的用于分析 AI/ML 模型切分的数据;

接收 NWDAF 发送的第十三请求结果;其中,所述第十三请求结果包括下述至少一项:待参与 AI/ML 模型切分的 UE(s)或者 SUPI、使用 AI/ML 模型的应用的标识、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量、AI/ML 模型的大小;若第十三请求中携带的参数包括推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息,则所述第十二请求结果中还包括:推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

将所述第十三请求结果作为所述第一消息。

41. 根据权利要求 32 或 33 所述的装置,其特征在于,接收单元,具体用于:

如果所述网络实体是新的网络实体 MMF,直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第十五请求,所述第十五请求用于请求 AI/ML 模型切分分析,所述第十五请求中携带的参数包括下述至少一项:与模型标识或模型切分标识关联的分析类型标识、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量;若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,所述第十五请求中携带的参数还包括:推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

根据所述第十五请求中携带的参数,确定 AF 与所述待参与模型联合推理的 UE(s)是否具备完全支持执行模型联合推理的能力;

若与所述待参与模型联合推理的 UE(s)具备完全支持执行模型联合推理的能力,则将所述第十五请求中携带的参数作为所述第一消息;

若与所述待参与模型联合推理的 UE(s)不具备完全支持执行模型联合推理的能力,则向 NWDAF 发送第十六请求,所述第十六请求中携带的参数包

括所述第十五请求中携带的参数，且所述第十六请求用于请求 NWDAF 通过 NF 查找其他可参与模型推理的 UE(s)以及其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量，若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第十六请求中携带的参数还包括：其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

接收 NWDAF 发送的第十六请求结果，并将所述第十六请求结果以及所述第十五请求中携带的参数作为所述第一消息；其中，所述第十六请求结果包括其他可参与模型推理的 UE(s)以及其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量，若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第十六请求结果还包括：其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息。

42. 根据权利要求 32 或 33 所述的装置，其特征在于，接收单元，具体用于：

如果所述网络实体是策略控制功能 PCF，直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第十七请求，所述第十七请求用于请求 AI/ML 模型切分分析；所述第十七请求中携带的参数包括下述至少一项：与模型标识或模型切分标识关联的分析类型标识、接收 AI/ML 模型切分的一个用户设备 UE 或一组 UEs 的标识或满足分析条件的任意 UEs、AI/ML 模型切分的区域、AI/ML 模型的大小、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量；若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第十七请求中携带的参数还包括：推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

根据所述第十七请求中携带的参数，向网络数据分析功能 NWDAF 发送第十八请求，所述第十八请求中携带的参数包括所述第十七请求中携带的参数，且所述第十八请求用于请求 NWDAF 向 5GC NF(s)采集 UE(s)对应的第五数据；所述第五数据中包括下述至少一项：待参与 AI/ML 模型切分的 UE(s)或者 SUPI、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推

理的 UE(s)可提供的内存、待参与模型联合推理的 UE(s)剩余的电量、AI/ML 模型的大小;若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,所述第五数据中还包括:推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

- 5 接收 NWDAF 发送的第十八请求结果;其中,所述第十八请求结果包括所述第五数据;

 将所述第十八请求结果作为所述第一消息。

43. 根据权利要求 32 或 33 所述的装置,其特征在于,接收单元,具体用于:

- 10 如果所述网络实体是策略控制功能 PCF,直接或通过网络能力开放功能 NEF 接收应用功能 AF 发送的第二十请求,所述第二十请求用于请求 AI/ML 模型切分分析,所述第二十请求中携带的参数包括下述至少一项:与模型标识或模型切分标识关联的分析类型标识、待参与模型联合推理的 UE(s)可提供的算力、待参与模型联合推理的 UE(s)可提供的内存、待参与模型联合推理的
15 UE(s)剩余的电量;若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,所述第二十请求中携带的参数还包括:推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

 根据所述第二十请求中携带的参数,确定所述 AF 以及待参与模型联合推理的 UE(s)是否具备完全支持执行模型联合推理的能力;

- 20 若所述 AF 以及待参与模型联合推理的 UE(s)具备完全支持执行模型联合推理的能力,则将所述第二十请求中携带的参数作为所述第一消息;

- 若所述 AF 以及待参与模型联合推理的 UE(s)不具备完全支持执行模型联合推理的能力,则向 NWDAF 发送第二十一请求,所述第二十一请求中携带的参数包括所述第二十请求中携带的参数,且所述第二十一请求用于请求
25 NWDAF 通过 NF 查找其他可参与模型推理的 UE(s)以及其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量,若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延,所述第二十一请求中携带的参数还包括:其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息;

- 30 接收 NWDAF 发送的第二十一请求结果,并将所述第二十一请求结果以

及所述第二十请求中携带的参数作为所述第一消息；其中，所述第二十一请求结果包括：其他可参与模型推理的 UE(s)以及其他可参与模型推理的 UE(s)可提供的算力、可提供的内存、剩余的电量，若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，所述第二十一请求结果还包括：其他可参与模型推理的 UE(s)推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息；

若第二十请求中携带的参数中不包含模型不同层的时延需求信息，则模型不同层的时延需求信息由 AF 提供。

44. 一种辅助模型切分的装置，其特征在于，应用于用户设备 UE，所述装置包括：

确定单元，用于根据自身能力信息，确定 AI/ML 模型切分点信息；

处理单元，用于将所述 AI/ML 模型切分点信息作为 AI/ML 模型切分的切分结果。

45. 根据权利要求 44 所述的装置，其特征在于，确定单元，具体用于：

根据自身能力信息中的可提供的算力、可提供的内存以及剩余的电量，确定模型切分点信息；

若推理 AI/ML 模型时需要数据集的隐私等级和/或模型不同层的时延，则根据所述自身能力信息中的可提供的算力、可提供的内存、剩余的电量、推理 AI/ML 模型需要的数据集的隐私等级和/或模型不同层的时延需求信息，确定模型切分点信息。

46. 一种处理器可读存储介质，其特征在于，所述处理器可读存储介质存储有计算机程序，所述计算机程序用于使所述处理器执行权利要求 1 至 29 任一项所述的方法。

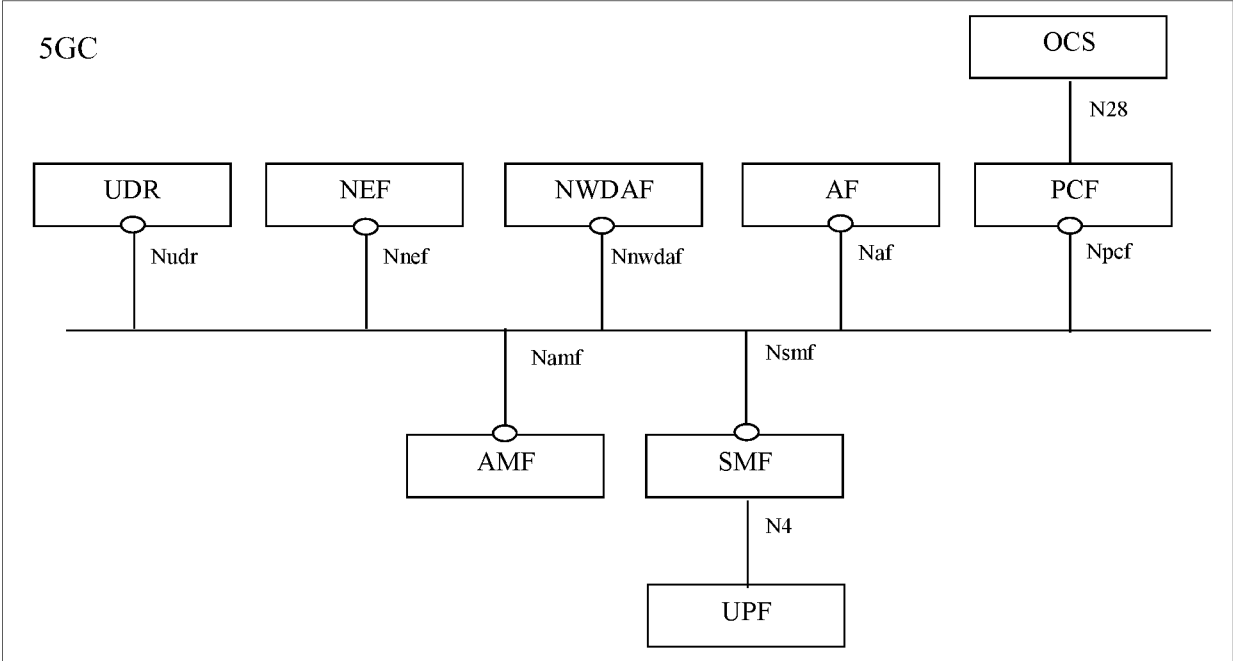


图 1

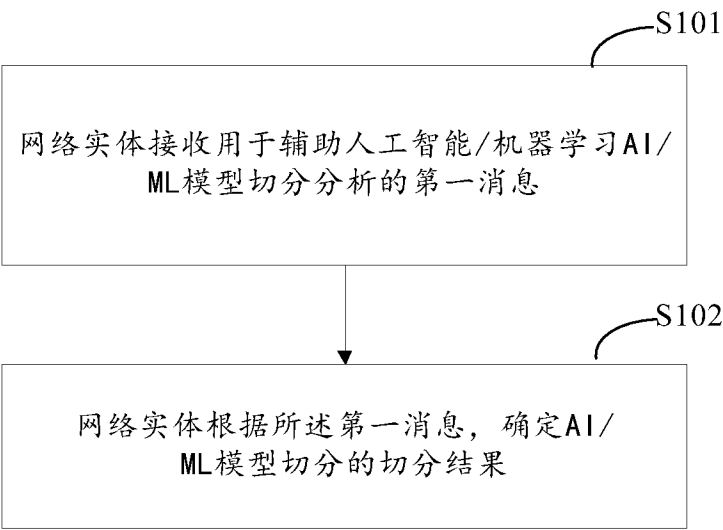


图 2

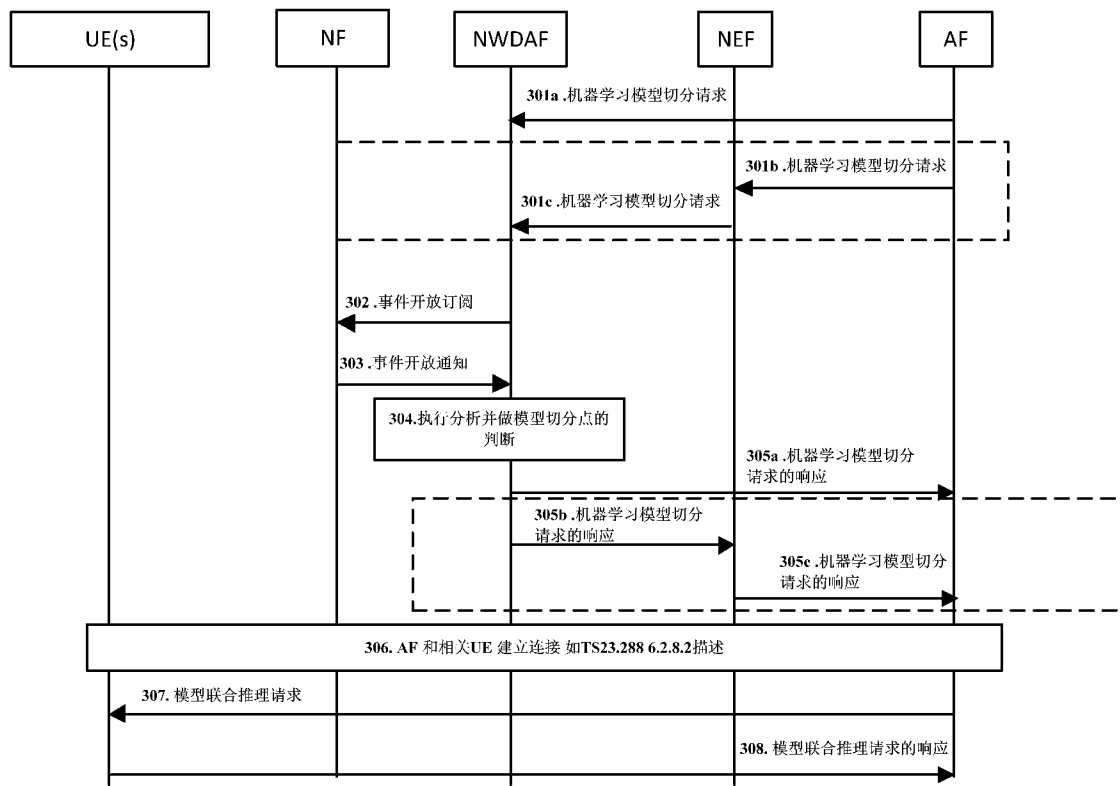


图 3

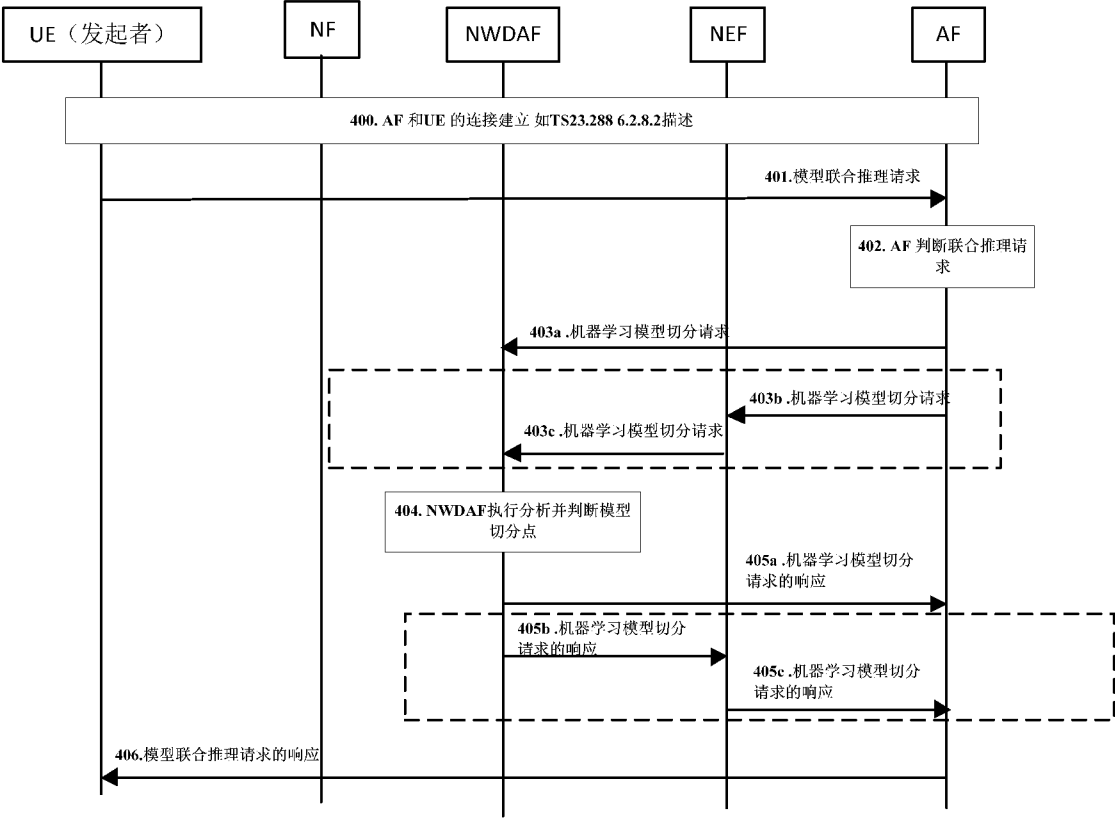


图 4

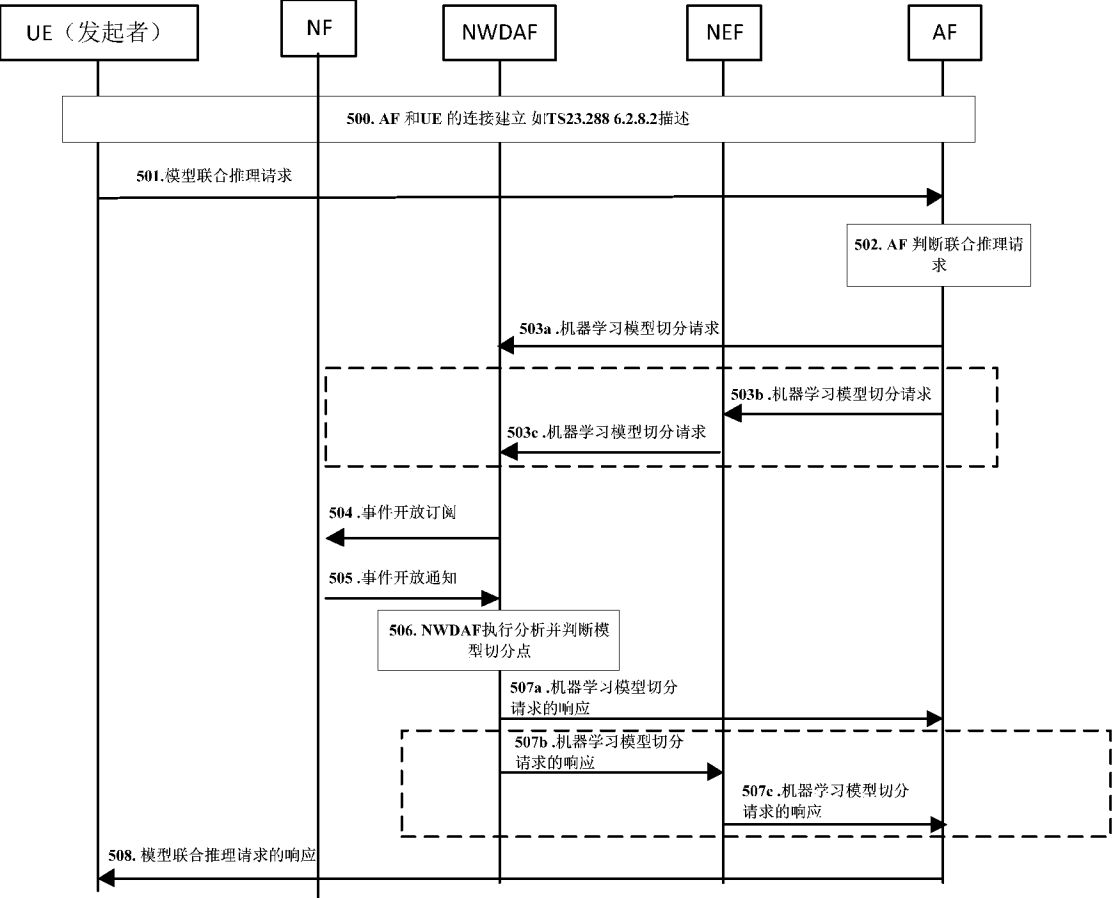


图 5

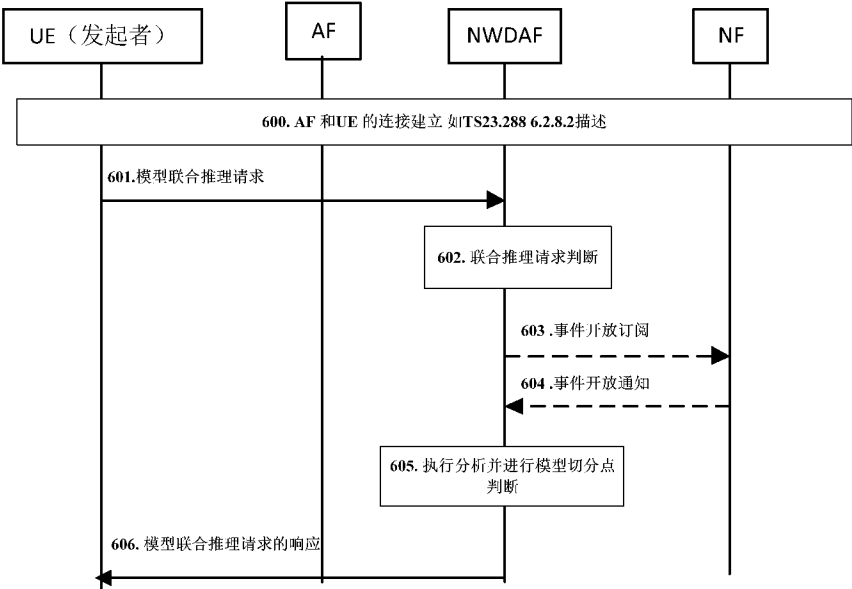


图 6

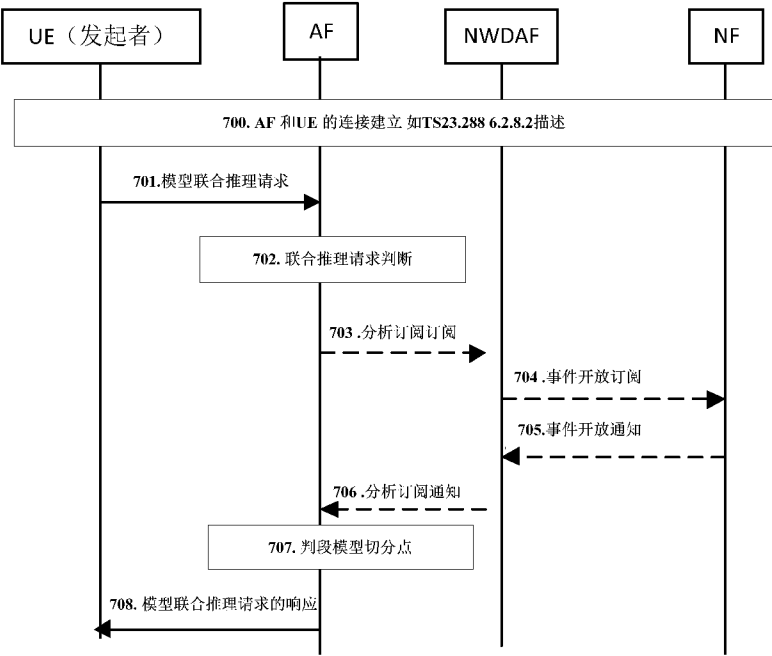


图 7

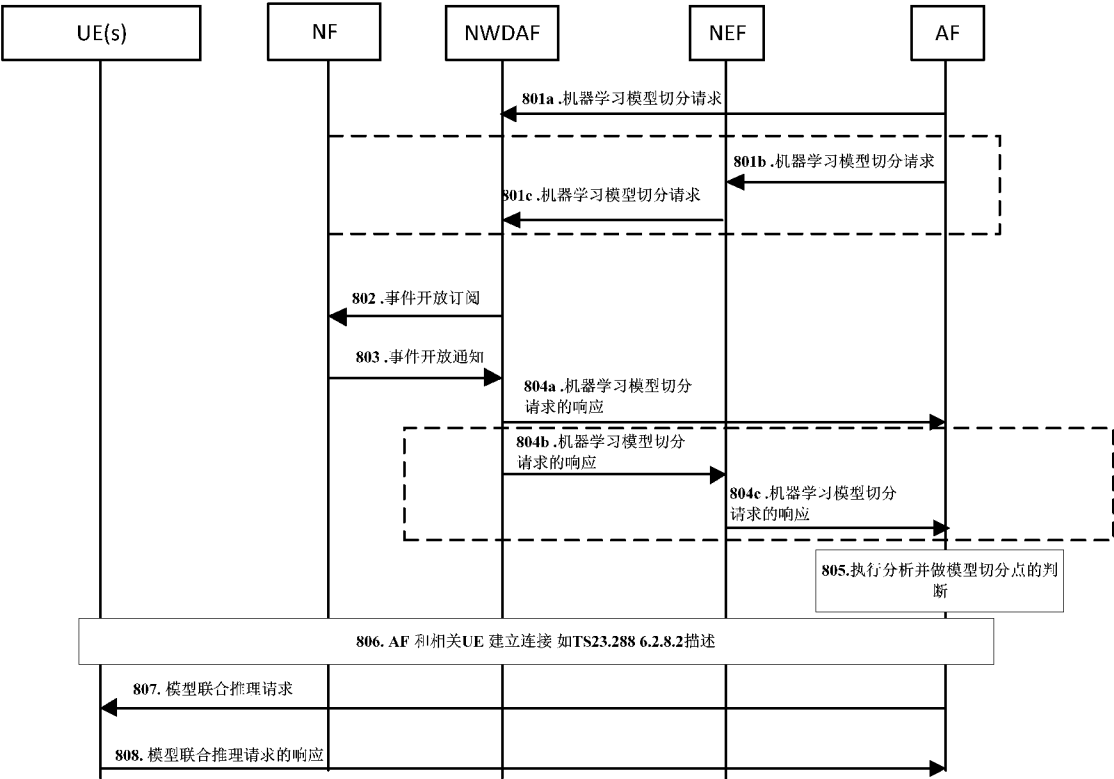


图 8

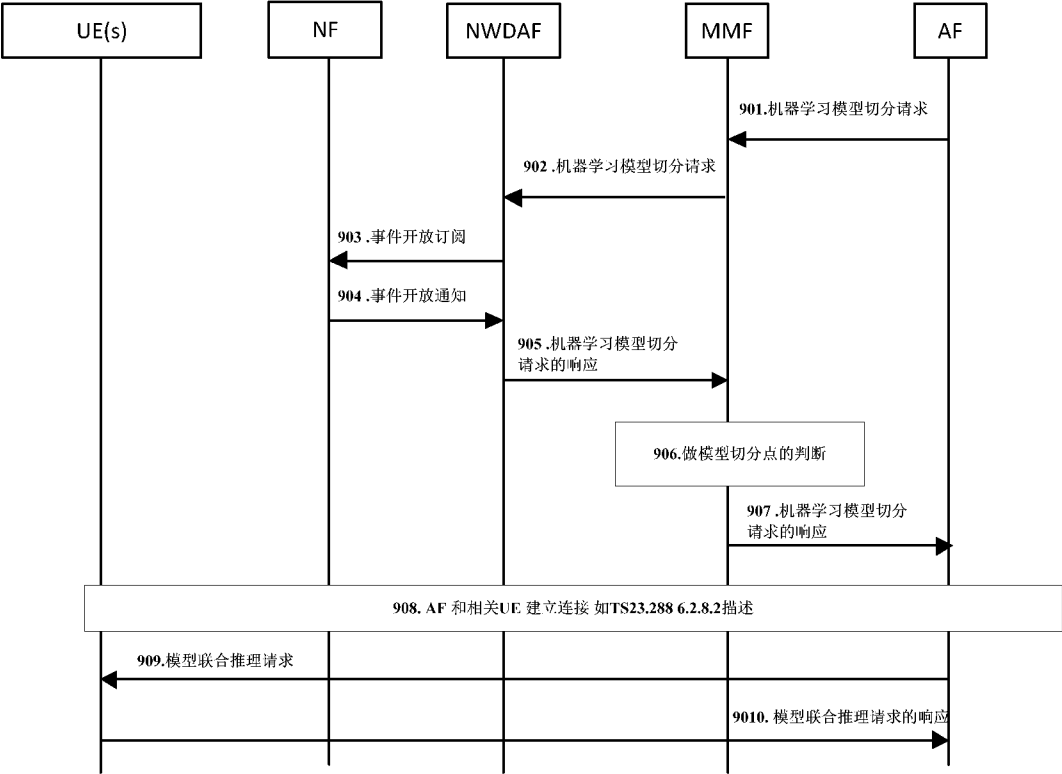


图 9

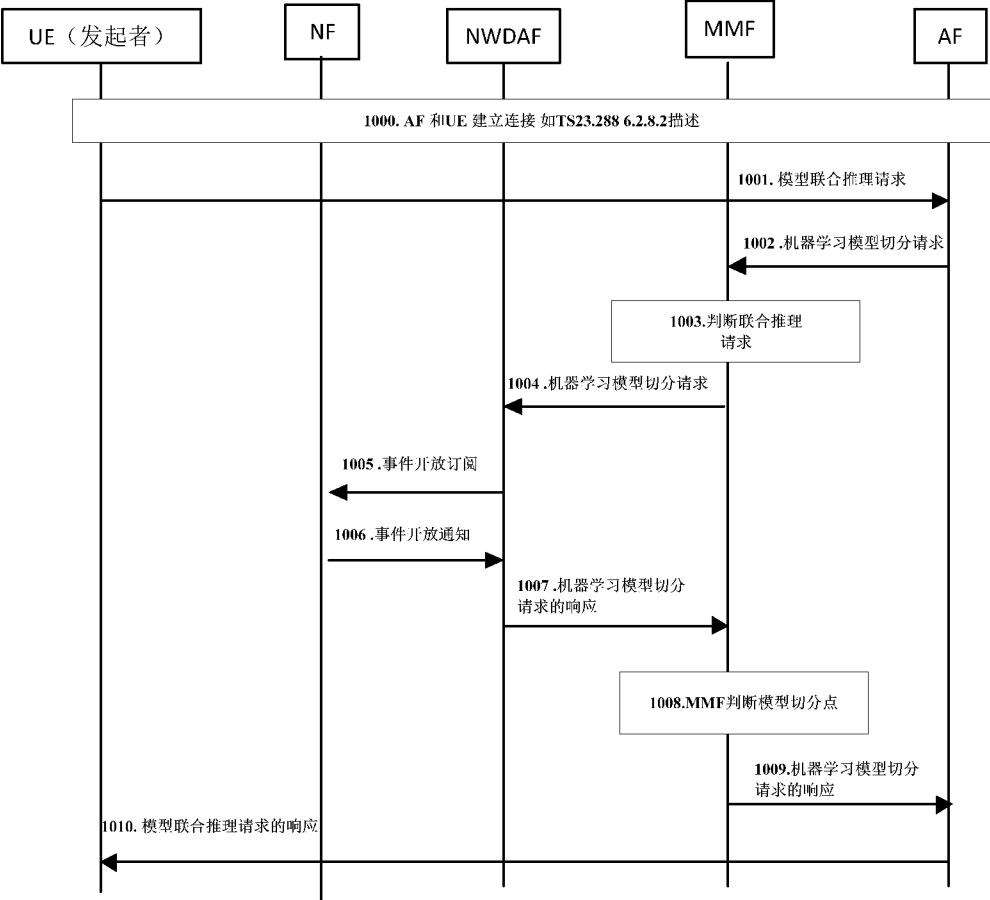


图 10

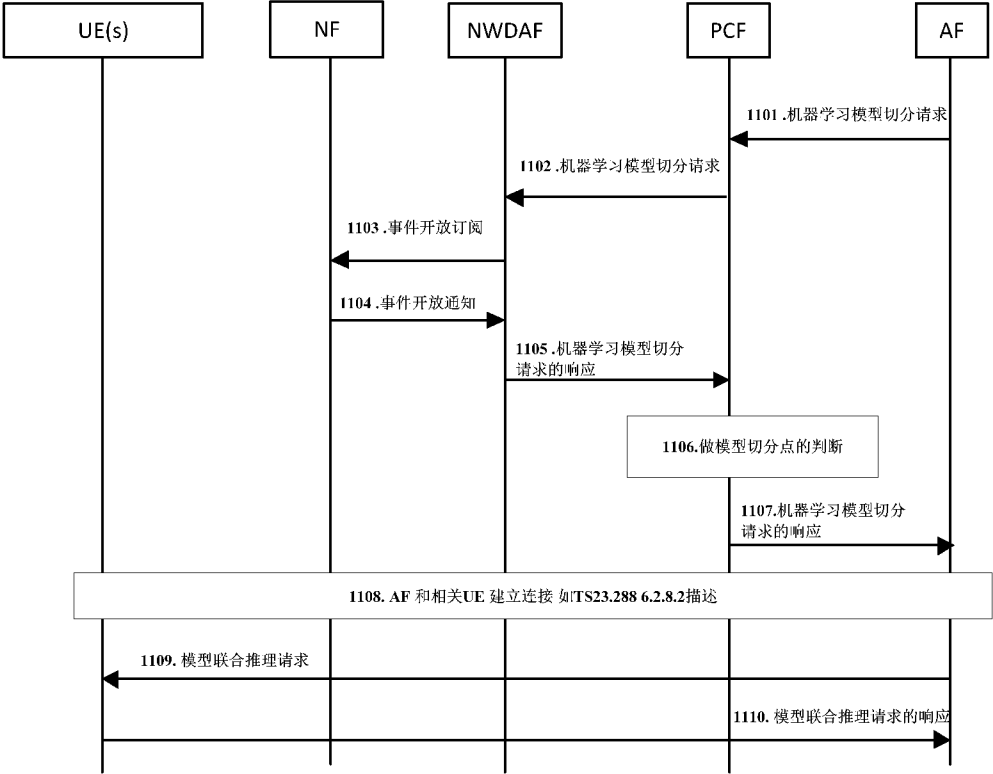


图 11

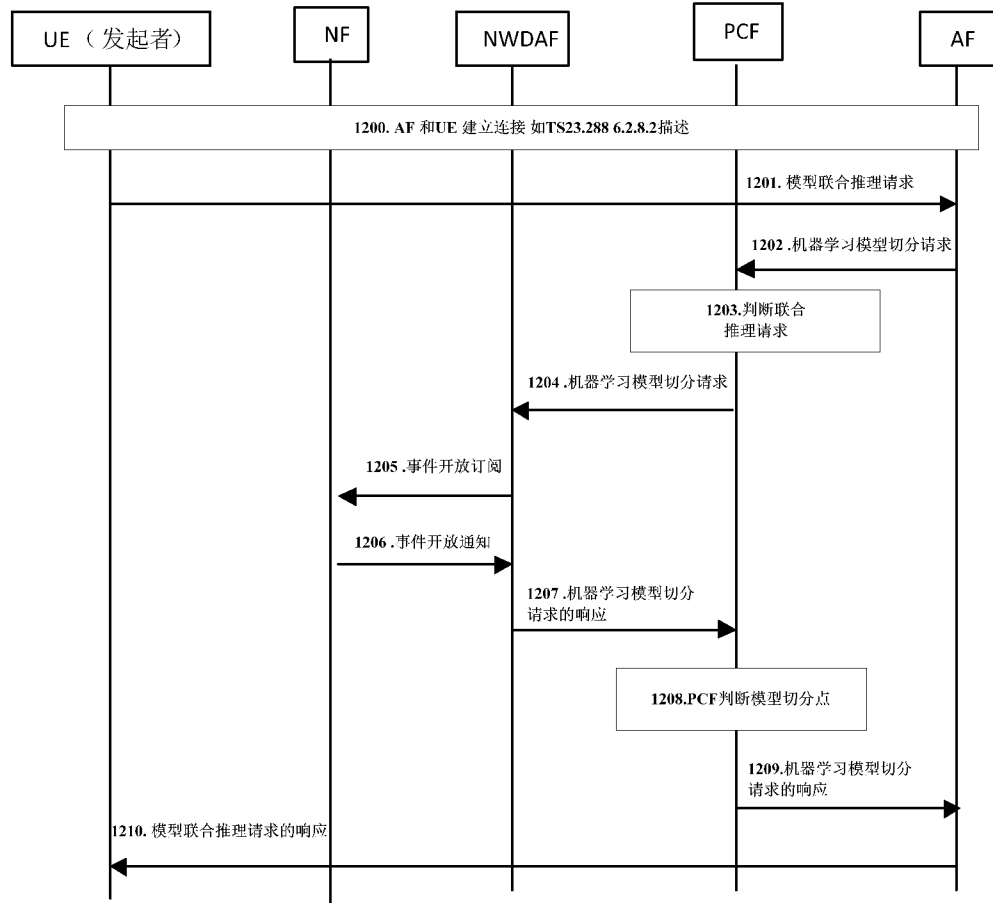


图 12

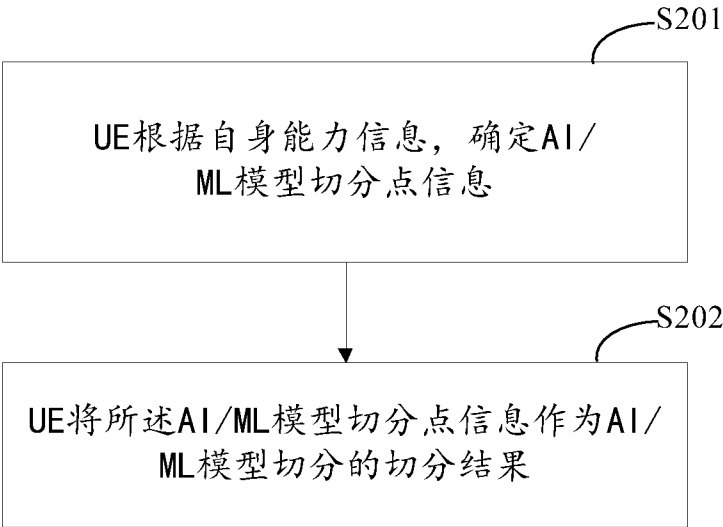


图 13

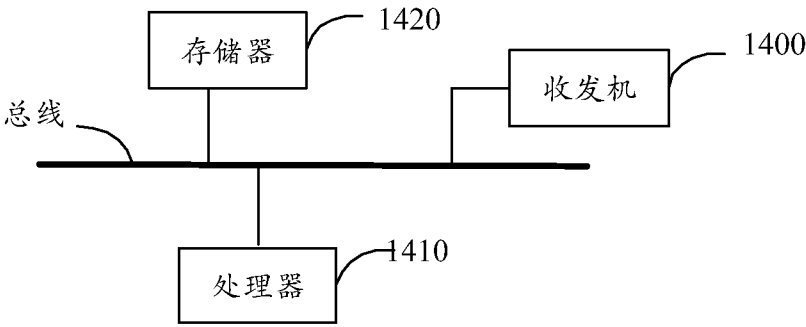


图 14

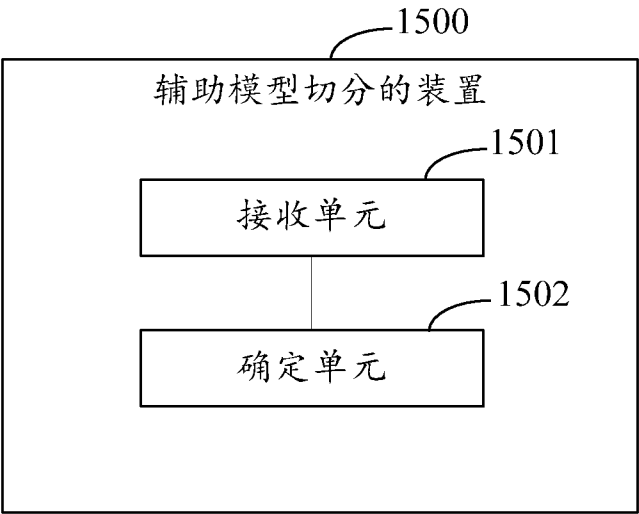


图 15

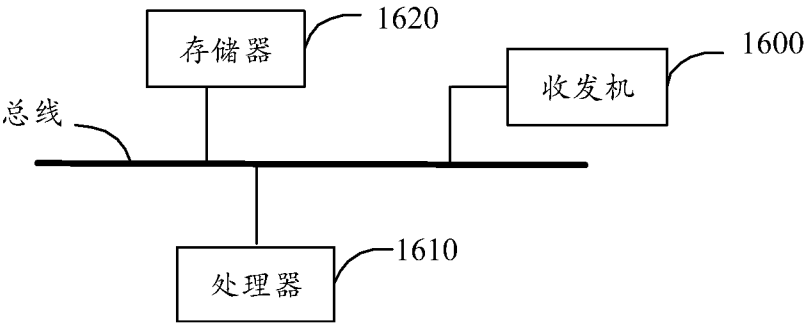


图 16

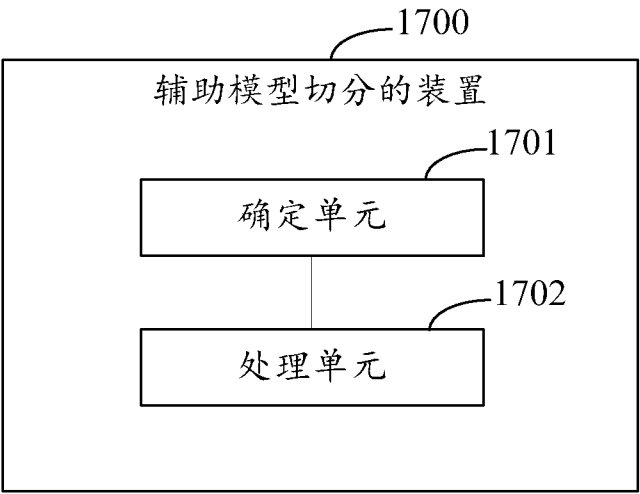


图 17

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2022/131852

A. CLASSIFICATION OF SUBJECT MATTER

G06N 20/00(2019.01)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)

G06N G06F

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)

CNABS, CNTXT, WOTXT, EPTXT, USTXT, CNKI, IEEE: 终端, 电量, 划分, 切分, 区分, 联合, 机器学习, 人工智能, 模型, 内存, 能力, 算力, 推理, 应用功能, 网络能力开放功能, 网络功能, 网络数据分析功能, 策略控制功能, UE, charge, division, slicing, union, ML, AI, model, memory, capacity, computational power, inference, 5G, AF, MMF, NEF, NF, NWDAF, PCF

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2021142637 A1 (GUANGDONG OPPO MOBILE TELECOMMUNICATIONS CORP., LTD.) 22 July 2021 (2021-07-22) description, page 1, paragraph 3-page 18, paragraph 7, and figures 1-20	1-3, 25-26, 30-33, 44-46
X	WO 2021163895 A1 (GUANGDONG OPPO MOBILE TELECOMMUNICATIONS CORP., LTD.) 26 August 2021 (2021-08-26) description, page 2, line 15-page 11, line 49, and figures 1-3	1-3, 25-26, 30-33, 44-46
A	CN 112287609 A (ZHEJIANG LAB) 29 January 2021 (2021-01-29) entire document	1-46
A	CN 113436208 A (INDUSTRIAL AND COMMERCIAL BANK OF CHINA CO., LTD.) 24 September 2021 (2021-09-24) entire document	1-46
A	zzh123666. "5G Standard Learning Note (3GPP TS 23.501)" https://blog.csdn.net/zzh123666/article/details/109706670 , 15 November 2020 (2020-11-15), pages 1-11	1-46



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

30 January 2023

Date of mailing of the international search report

07 February 2023

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

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2022/131852

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
WO	2021142637	A1	22 July 2021	EP	4087213	A1	09 November 2022
				US	2022334881	A1	20 October 2022
WO	2021163895	A1	26 August 2021	None			
CN	112287609	A	29 January 2021	None			
CN	113436208	A	24 September 2021	None			

国际检索报告

国际申请号

PCT/CN2022/131852

A. 主题的分类

G06N 20/00 (2019.01) 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)

G06N G06F

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))

CNABS, CNTXT, WOTXT, EPTXT, USTXT, CNKI, IEEE: 终端, 电量, 划分, 切分, 区分, 联合, 机器学习, 人工智能, 模型, 内存, 能力, 算力, 推理, 应用功能, 网络能力开放功能, 网络功能, 网络数据分析功能, 策略控制功能, UE, charge, division, slicing, union, ML, AI, model, memory, capacity, computational power, inference, 5G, AF, MMF, NEF, NF, NWDAF, PCF

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 2021142637 A1 (GUANGDONG OPPO MOBILE TELECOMMUNICATIONS CORP., LTD.) 2021 年7月22日 (2021 - 07 - 22) 说明书第1页第3段至第18页第7段, 附图1-20	1-3, 25-26, 30-33, 44-46
X	WO 2021163895 A1 (GUANGDONG OPPO MOBILE TELECOMMUNICATIONS CORP., LTD.) 2021 年8月26日 (2021 - 08 - 26) 说明书第2页第15行至第11页49行, 附图1-3	1-3, 25-26, 30-33, 44-46
A	CN 112287609 A (之江实验室) 2021年1月29日 (2021 - 01 - 29) 全文	1-46
A	CN 113436208 A (中国工商银行股份有限公司) 2021年9月24日 (2021 - 09 - 24) 全文	1-46
A	zzh123666. "5G标准学习笔记 (3GPP TS 23.501)" https://blog.csdn.net/zzh123666/article/details/109706670 , 2020年11月15日 (2020 - 11 - 15), 第1-11页	1-46

☐ 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

2023年1月30日

国际检索报告邮寄日期

2023年2月7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

传真号 (86-10)62019451

授权官员

徐淑娴

电话号码 (+86) 010-53961535

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2022/131852

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	2021142637	A1	2021年7月22日	EP	4087213	A1	2022年11月9日
				US	2022334881	A1	2022年10月20日
WO	2021163895	A1	2021年8月26日	无			
CN	112287609	A	2021年1月29日	无			
CN	113436208	A	2021年9月24日	无			