(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112488563 B
(45)授权公告日 2023.06.06

(21)申请号 202011460037.4

(22)申请日 2020.12.11

(65)同一申请的已公布的文献号
申请公布号 CN 112488563 A

(43)申请公布日 2021.03.12

(73)专利权人 中国联合网络通信集团有限公司
地址 100033 北京市西城区金融大街21号

(72)发明人 李建飞 曹畅 何涛 李铭轩

(74)专利代理机构 北京中博世达专利商标代理有限公司 11274
专利代理师 申健

(51)Int.Cl.
G06Q 10/0639(2023.01)
G06N 3/0464(2023.01)

(56)对比文件
CN 110532092 A,2019.12.03
CN 111131516 A,2020.05.08
CN 111401516 A,2020.07.10
CN 111866775 A,2020.10.30
CN 111818585 A,2020.10.23
CN 110659134 A,2020.01.07
CN 110782343 A,2020.02.11
CN 112035247 A,2020.12.04
CN 111126594 A,2020.05.08
CN 111144561 A,2020.05.12
US 2019050388 A1,2019.02.14
李建飞 等.算力网络中面向业务体验的算力建模.《中兴通讯技术》.2020,第26卷(第5期),第34-38,52页.
Zhi Yang 等.A Cost-Based Resource Scheduling Paradigm in Cloud Computing.《2011 12th International Conference on Parallel and Distributed Computing, Applications and Technologies》.2012,第417-422页.
Haifeng Lu 等.Optimization of lightweight task offloading strategy for mobile edge computing based on deep reinforcement learning.《Future Generation Computer Systems》.2019,第102卷第847-861页.

审查员 王黎

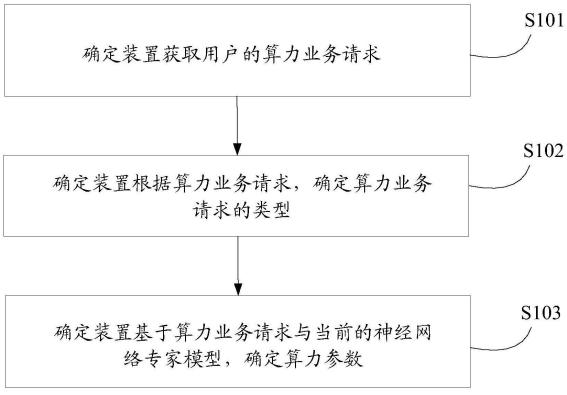
权利要求书2页 说明书13页 附图4页

(54)发明名称

一种算力参数的确定方法和装置

(57)摘要

本申请公开了一种算力参数的确定方法和装置,涉及网络通信领域,用于提高确定算力参数的效率。该确定方法包括:获取用户的算力业务请求;根据算力业务请求,确定算力业务请求的参数类型;基于算力业务请求与当前的神经网络模型,确定算力参数;当前的神经网络模型与算力业务请求的参数类型相对应。相比于现有技术中人工确定业务请求对应的算力参数的方法,需要用户具备一定的专业能力,而且效率较低,本申请提供的算力参数的确定方法是基于神经网络模型,确定算力参数,从而不需要用户具备一定的专业能力,进而能够提高确定算力参数的效率。



CN 112488563 B

1.一种算力参数的确定方法,其特征在于,所述确定方法包括:

获取用户的算力业务请求;

根据所述算力业务请求,确定所述算力业务请求的类型;所述算力业务请求包括以下至少之一:算力参数、关键算力参数、业务需求参数、关键业务需求参数、抽象需求参数;所述算力业务请求的类型包括以下至少之一:算力参数类型、业务需求类型、以及抽象需求类型;所述关键算力参数为满足所述算力业务的算力参数;所述关键业务需求参数为满足所述算力业务的业务需求参数;

其中,在所述算力参数的数量大于或等于第一预设阈值;或者所述关键算力参数的数量大于或等于第二预设阈值的情况下,确定所述算力业务请求的类型为所述算力参数类型;

在所述业务需求参数的数量大于或等于第三预设阈值;或者所述关键业务需求参数的数量大于或等于第四预设阈值的情况下,确定所述算力业务请求的类型为所述业务需求类型;

在所述算力参数的数量小于所述第一预设阈值,所述业务需求参数的数量小于所述第二预设阈值,以及所述算力业务请求中包括所述抽象需求参数的情况下,确定所述算力业务请求的类型为所述抽象需求类型;

基于所述算力业务请求与当前的神经网络模型,确定算力参数;所述当前的神经网络模型与所述算力业务请求的类型相对应。

2.根据权利要求1所述的确定方法,其特征在于,在所述算力参数的数量大于或等于所述第一预设阈值,所述业务需求参数的数量大于或等于所述第三预设阈值,以及所述算力业务请求中包括所述算力参数、所述业务需求参数、以及所述抽象需求参数的情况下,确定所述算力业务请求的类型为所述算力参数类型、所述业务需求类型、以及所述抽象需求类型中任意一种类型。

3.根据权利要求2所述的确定方法,其特征在于,所述确定所述算力业务请求的类型,包括:

根据所述算力业务请求中所包含的关键字,确定所述算力业务请求的类型。

4.根据权利要求1所述的确定方法,其特征在于,所述方法还包括:

确定所述算力参数与目标算力参数的差值小于或等于预设阈值;所述目标算力参数为满足所述算力业务请求的算力参数;

基于所述算力参数和所述算力业务请求,优化所述当前的神经网络模型。

5.一种算力参数的确定装置,其特征在于,所述确定装置包括:

获取单元,用于获取用户的算力业务请求;

确定单元,用于根据所述获取单元获取的所述算力业务请求,确定所述算力业务请求的类型;所述算力业务请求包括以下至少之一:算力参数、关键算力参数、业务需求参数、关键业务需求参数、抽象需求参数;所述算力业务请求的类型包括以下至少之一:算力参数类型、业务需求类型、以及抽象需求类型;所述关键算力参数为满足所述算力业务的算力参数;所述关键业务需求参数为满足所述算力业务的业务需求参数;

其中,在所述算力参数的数量大于或等于第一预设阈值;或者所述关键算力参数的数量大于或等于第二预设阈值的情况下,确定所述算力业务请求的类型为所述算力参数类

型；

在所述业务需求参数的数量大于或等于第三预设阈值；或者所述关键业务需求参数的数量大于或等于第四预设阈值的情况下，确定所述算力业务请求的类型为所述业务需求类型；

在所述算力参数的数量小于所述第一预设阈值，所述业务需求参数的数量小于所述第二预设阈值，以及所述算力业务请求中包括所述抽象需求参数的情况下，确定所述算力业务请求的类型为所述抽象需求类型；所述确定单元，还用于基于所述获取单元获取的所述算力业务请求与当前的神经网络模型，确定算力参数；所述当前的神经网络模型与所述算力业务请求的类型相对应。

6.根据权利要求5所述的确定装置，其特征在于，在所述算力参数的数量大于或等于所述第一预设阈值，所述业务需求参数的数量大于或等于所述第三预设阈值，以及所述算力业务请求中包括所述算力参数、所述业务需求参数、以及所述抽象需求参数的情况下，确定所述算力业务请求的类型为所述算力参数类型、所述业务需求类型、以及所述抽象需求类型中任意一种类型。

7.根据权利要求6所述的确定装置，其特征在于，所述确定单元具体用于：

根据所述算力业务请求包含的关键字，确定所述算力业务请求的类型。

8.根据权利要求5所述的确定装置，其特征在于，所述确定单元还用于：

确定所述算力参数与目标算力参数的差值小于或等于预设阈值；所述目标算力参数为满足所述算力业务请求的算力参数；

基于所述算力参数和所述算力业务请求，优化所述当前的神经网络模型。

9.一种算力参数的确定设备，其特征在于，所述算力参数的确定设备包括存储器和处理器；所述存储器和所述处理器耦合；所述存储器用于存储计算机程序代码，所述计算机程序代码包括计算机指令；当所述处理器执行所述计算机指令时，所述算力参数的确定设备执行如权利要求1-4中任意一项所述的算力参数的确定方法。

10.一种计算机可读存储介质，其特征在于，所述计算机可读存储介质中存储有指令，当所述计算机可读存储介质在算力参数的确定设备上运行时，使得所述设备执行权利要求1-4任一项所述的算力参数的确定方法。

一种算力参数的确定方法和装置

技术领域

[0001] 本申请涉及网络通信领域，尤其涉及一种算力参数的确定方法和装置。

背景技术

[0002] 近年来，人工智能(artificial intelligence，AI)已经成为当代社会一项通用的技术，也是将来智能社会必不可少的技术。算力，算法和数据是AI的三要素。算力作为AI的基础，直接影响着AI业务的应用与部署。目前已经兴起的AI行业以及潜在的智能业务都对算力提出了较高的要求。

[0003] 不同的业务请求对应的算力参数不同，而且业务请求的内容和形式多种多样。因此，在业务请求的描述比较抽象时，需要用户人工分析业务请求的内容，从而确定业务请求对应的算力参数。但是，该人工确定业务请求对应的算力参数的方法，需要用户具备一定的专业能力，而且效率较低。

发明内容

[0004] 本申请提供了一种算力参数的确定方法和装置，用于提高确定算力参数的效率。

[0005] 为达到上述目的，本申请采用如下技术方案：

[0006] 第一方面，本申请提供了一种算力参数的确定方法。该确定方法包括：获取用户的算力业务请求。之后，根据该算力业务请求，确定该算力业务请求的类型，并基于该算力业务请求与该算力业务请求的类型相对应的当前的神经网络模型，确定目标算力参数。

[0007] 本申请提供的算力参数的确定方法，通过获取用户的算力业务请求，基于算力业务请求与当前的神经网络模型，确定算力参数。相比于现有技术中人工确定业务请求对应的算力参数的方法，需要用户具备一定的专业能力，而且效率较低，本申请提供的算力参数的确定方法是基于神经网络模型，确定算力参数，从而不需要用户具备一定的专业能力，进而能够提高确定算力参数的效率。

[0008] 第二方面，本申请提供了一种算力参数的确定装置。该装置包括：获取单元，获取用户的算力业务请求；确定单元，用于根据上述获取单元获取的算力业务请求，确定该算力业务请求的类型；该确定单元，还用于基于上述获取单元获取的算力业务请求与当前的神经网络模型，确定算力参数；该当前的神经网络模型与上述算力业务请求的类型相对应。

[0009] 第三方面，本申请提供一种算力参数的确定设备，该算力参数的确定设备包括存储器和处理器。存储器和处理器耦合。该存储器用于存储计算机程序代码，该计算机程序代码包括计算机指令。当处理器执行计算机指令时，该算力参数的确定设备执行如第一方面及其任一种可能的设计方式所述的算力参数的确定方法。

[0010] 第四方面，本申请提供了一种计算机可读存储介质，计算机可读存储介质中存储有指令，当所述计算机可读存储介质在算力参数的确定装置上运行时，使得该装置执行如第一方面及其任一种可能的设计方式所述算力参数的确定方法。

[0011] 第五方面，本申请提供一种计算机程序产品，该计算机程序产品包括计算机指令，

当所述计算机指令在算力参数的确定装置上运行时，使得所述算力参数的确定装置执行如第一方面及其任一种可能的设计方式所述的算力参数的确定方法。

[0012] 本申请中第二方面到第五方面及其各种实现方式的具体描述，可以参考第一方面及其各种实现方式中的详细描述；并且，第二方面到第五方面及其各种实现方式的有益效果，可以参考第一方面及其各种实现方式中的有益效果分析，此处不再赘述。

[0013] 本申请的这些方面或其他方面在以下的描述中会更加简明易懂。

附图说明

[0014] 为了更清楚地说明本申请实施例中的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0015] 图1为本申请提供的算力参数的确定方法的流程示意图一；

[0016] 图2为本申请提供的第一神经网络专家系统的结构示意图；

[0017] 图3为本申请提供的第二神经网络专家系统的结构示意图；

[0018] 图4为本申请提供的算力参数的确定方法的流程示意图二；

[0019] 图5为本申请实施例提供的算力参数的确定设备的硬件结构示意图；

[0020] 图6为本申请实施例提供的算力参数的确定装置的结构示意图。

具体实施方式

[0021] 下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

[0022] 术语“第一”、“第二”仅用于描述目的，而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”的特征可以明示或者隐含地包括一个或者更多个该特征。在本申请的描述中，除非另有说明，“多个”的含义是两个或两个以上。

[0023] 近年来，人工智能(artificial intelligence，AI)已经成为当代社会一项通用的技术，也是将来智能社会必不可少的技术。算力，算法和数据是AI的三要素。算力作为AI的基础，直接影响着AI业务的应用与部署。目前已经兴起的AI行业以及潜在的智能业务都对算力提出了较高的要求。

[0024] 不同的业务请求对应的算力参数不同，而且业务请求的内容和形式多种多样。因此，在业务请求的描述比较抽象时，需要用户人工分析业务请求的内容，从而确定业务请求对应的算力参数。但是，该人工确定业务请求对应的算力参数的方法，需要用户具备一定的专业能力，而且效率较低。

[0025] 针对上述问题，本申请提供了一种算力参数的确定方法，通过获取用户的算力业务请求，基于算力业务请求与当前的神经网络模型，确定算力参数。相比于现有技术中人工确定业务请求对应的算力参数的方法，需要用户具备一定的专业能力，而且效率较低，本申

请提供的算力参数的确定方法是基于神经网络模型，确定算力参数，从而不需要用户具备一定的专业能力，进而能够提高确定算力参数的效率。

[0026] 本申请实施例提供的算力参数的确定方法的执行主体为算力参数的确定装置(后续简称为:确定装置)。该确定装置可以与算力网络系统集成在一起，也可以独立设置，本申请实施例对此不作限定。

[0027] 算力网络系统是一种能够将当前的计算能力状况和网络状况作为路由信息发布到网络，网络将计算任务报文路由到相应的计算节点，实现用户体验最优、计算资源利用率最优、网络效率最优的网络系统。通过算力网络系统内建计算任务动态路由的能力，根据业务需求，基于实时的计算资源性能、网络性能、成本等多维因素，动态、灵活地调度计算任务，从而提高资源利用率，网络利用效率，提高业务用户体验。面向边缘计算场景，可通过算力网络系统实现边缘计算成网，实现边边协作，利用服务的多实例、多副本特性，实现用户的就近接入和服务的负载均衡，以解决其部署复杂、效率低、资源复用率低等问题，助力边缘计算规模部署。

[0028] 下面对本申请实施例提供的算力参数的确定方法进行描述。

[0029] 如图1所示，该算力参数的确定方法包括：

[0030] S101、确定装置获取用户的算力业务请求。

[0031] 可选的，确定装置可以接收用户输入的算力业务请求。

[0032] 可选的，算力业务请求可以包括渲染业务、训练任务、推理业务、视频播放、超算类或其他业务。

[0033] 算力业务请求包括算力参数、用户的业务需求参数或者用户的抽象需求参数。

[0034] 算力参数为算力网络系统满足算力业务所需要的底层硬件资源所对应的参数。

[0035] 可选的，算力参数可以包括中央处理器(central processing unit,CPU)的参数，图形处理器(graphics processing unit,GPU)的参数，存储的参数，网络指标的参数等。

[0036] 用户的业务需求参数为用户对算力业务的具体需求。

[0037] 示例性的，用户的业务需求参数可以包括用户对业务的需求参数、对模型镜像的需求参数、对算法的需求参数、对时间的需求参数、对运行次数的需求参数。

[0038] 用户的抽象需求参数为用户对算力业务请求的抽象描述。

[0039] 示例性的，用户的抽象需求参数为训练一个人脸识别的神经网络模型，或，用户的抽象需求参数为对一部电影做后续渲染。

[0040] S102、确定装置根据算力业务请求，确定算力业务请求的类型。

[0041] 算力业务请求的类型包括算力参数类型、业务需求类型或抽象需求类型。

[0042] 可选的，确定装置可以根据算力业务请求中所包含的关键字，确定算力业务请求的类型。

[0043] 关键字包括算力参数关键字、业务需求关键字以及抽象需求关键字。

[0044] 可选的，算力参数关键字包括CPU、GPU、张量运算、存储、编解码、网络指标。

[0045] 例如，如表1所示，在算力业务请求中包含关键字CPU时，确定CPU的参数10每秒万亿次操作(tera operations per second,TOPS)为算力参数；在算力业务请求中包含关键字GPU时，确定GPU的参数12(floating point operations per second,TFLOPs)为算力参数；在算力业务请求中包含关键字张量运算时，确定张量运算嵌入式神经网络处理器

(neural-networ k processing unit,NPU)/张量处理单元(tensor processing unit,TPU)的参数8TFLOPS为算力参数;在算力业务请求中包含关键字存储时,确定存储的参数10千兆为算力参数;在算力业务请求中包含关键字网络指标时,确定网络指标的参数时延小于50毫秒(millisecond,ms)为算力参数;在算力业务请求中包含关键字编解码时,确定编解码为满足标准为算力参数。

[0046] 表1

[0047]

逻辑运算CPU	矩阵运行GPU	张量运算NPU/TPU	存储大小	编解码需求	网络指标
10TOPS	12TFLOPs	8TFLOPS	10G	满足标准	时延<50 ms

[0048] 在算力业务请求中包括的算力参数的个数大于或等于第一预设阈值,确定算力业务请求的类型为算力参数类型。

[0049] 第一预设阈值可以根据实际情况确定,本申请对此并不进行限定。

[0050] 进一步的,若算力业务请求中的关键算力参数的数量大于或等于第二预设阈值,则确定算力业务请求的类型为算力参数类型。

[0051] 第二预设阈值可以根据实际情况确定,本申请对此并不进行限定。

[0052] 关键算力参数为算力参数中满足算力业务所必须的算力参数。

[0053] 示例性的,关键算力参数可以为CPU的参数,GPU的参数,存储的参数,网络指标的参数。

[0054] 通过关键算力参数来判断算力业务请求的类型为算力参数类型,可以提高确定算力业务请求的类型为算力参数类型的准确性。

[0055] 可选的,业务需求关键字包括业务、模型镜像、算法、时间、运行次数以及数据量。

[0056] 例如,如表2所示,在确定算力业务请求中包含关键字业务时,确定对业务的需求参数视频渲染或模型训练为业务需求参数;在确定算力业务请求中包含关键字模型镜像时,确定对模型镜像的需求参数光线追踪与全域光渲染KeyShot或残差网络ResNet为业务需求参数;在确定算力业务请求中包含关键字算法时,确定对算法的需求参数分辨率1080逐行扫描(progressiv e scanning,p)或152层网络和6千个参数为业务需求参数;在确定算力业务请求中包含关键字时间时,确定对时间的需求参数一个月或一周为业务需求参数;在确定算力业务请求中包含关键字运行次数时,确定对运行次数的需求参数一次或十万次为业务需求参数;在确定算力业务请求中包含关键字数据量时,确定对数据量的需求参数10G视频或10G视频为业务需求参数。

[0057] 表2

[0058]

业务	模型镜像	算法参数	时间要求	运行次数	数据量

	视频渲染	KeyShot	分辨率 1080p	1 个月	1 次	10G 视频
[0059]	模型训练	ResNet	152 层网络，6k 个参数	一周	10w 次	100G 图片

[0060] 在确定算力业务请求中包括的业务需求参数的个数大于或等于第三预设阈值时，确定算力业务请求的类型为业务需求类型。

[0061] 第三预设阈值可以根据实际情况确定，本申请对此并不进行限定。

[0062] 进一步的，若算力业务请求中的关键业务需求参数的数量大于或等于第四预设阈值，则确定算力业务请求的类型为业务需求参数类型。

[0063] 第四预设阈值可以根据实际情况确定，本申请对此并不进行限定。

[0064] 关键业务需求参数为算力参数中满足算力业务所必须的业务需求参数。

[0065] 关键业务需求参数可以为对业务的需求参数、对模型镜像的需求参数、对算法的需求参数、对时间的需求、对运行次数的需求。

[0066] 通过关键业务需求参数来判断算力业务请求的类型为业务需求参数类型，可以提高确定算力业务请求的类型为业务需求参数类型的准确性。

[0067] 在确定算力业务请求中所包括的算力参数的个数小于第一预设阈值，业务需求参数的个数小于第二预设阈值，且包括抽象需求参数时，确定算力业务请求的类型为抽象需求类型。

[0068] 抽象需求参数即为用户对算力业务请求的抽象描述。

[0069] 示例性的，用户的算力业务请求为训练一个人脸识别的神经网络模型，算力业务请求中不包括算力参数，也不包括业务需求参数，而且训练一个人脸识别的神经网络模型为抽象需求参数，则算力业务请求的类型为抽象需求类型。

[0070] 需要说明的是，在算力业务请求中包括算力参数、业务需求参数以及抽象需求参数，且算力参数的个数大于或等于第一预设阈值、业务需求参数的个数大于或等于第三预设阈值时，算力业务请求的类型可以为算力参数类型，也可以为业务需求类型，还可以为抽象需求类型。

[0071] 可选的，确定装置在获取用户的算力业务请求的类型为算力参数类型时，确定装置确定算力业务请求中所包含的算力参数为目标算力参数。在算力参数的类型为业务需求类型或抽象需求类型的情况下，继续执行S103。

[0072] S103、确定装置基于算力业务请求与当前的神经网络模型，确定算力参数。

[0073] 神经网络模型包括第一神经网络模型和第二神经网络模型。

[0074] 第一神经网络模型与算力业务请求的类型为业务需求类型相对应。即在算力业务请求的类型为业务需求类型时，当前的神经网络模型为第一神经网络模型。

[0075] 第一神经网络模型用于根据业务需求参数预测算力参数。

[0076] 可选的，第一神经网络模型可以为第一神经网络专家系统中的神经网络模型。

[0077] 如图2所示，第一神经网络专家系统20包括第一神经网络模型21、第一规则库22以

及第一输出解释模块23。

[0078] 第一神经网络专家系统20的输入为算力业务请求中的业务需求参数，第一神经网络专家系统20的输出为算力参数。

[0079] 第一规则库22用于保存业务需求参数与算力参数对应的专家知识，并对第一神经网络模型21进行训练。

[0080] 可选的，第一规则库22可以保存从实际经验中获取的算力参数与业务需求参数对应的专家知识。专家知识可以是来自实际运行的大量实例集，也可以是从领域中的专家或文献资料中获取的大量规则，或者是二者的混合。

[0081] 示例性的，如表3所示，第一规则库22包括业务需求参数为推荐经典AlexNet模型配置、80w节点、6096w参数、卷积层参数3.8％、全连接层参数96.2％、推荐存储为10G、推荐完成时间为1周、推荐网络配置为非实时，对应的算力参数为逻辑运算为10TOPS、矩阵运行为12TFLOPS、张量运算为8TFLOPS、存储为10G、网络为非实时；业务需求参数为8路视频、清晰度为1080p、存储为2T、网络为延迟<200ms、上行带宽为xxMB、下行带宽为xxMb、编解码能力为xxx标准，对应的算力参数为逻辑运算为3TO PS、矩阵运行为5TFLOPS、存储为25T、网络为延迟<200ms、上行带宽为xxMB、下行带宽为xxMb、编解码为2个解码引擎和4个编码引擎。

[0082] 表3

[0083]

序号	业务需求参数	算力参数
1	计算能力: 推荐经典 AlexNet 模型配置: 80w 节点， 6096w 参数， 卷积层参数 3.8%	计算单元 逻辑运算: 10TOPS 矩阵运行: 12TFLOPS 张量运算: 8TFLOPS ------存储------- 10G

[0084]

	全连接层参数 96.2% ---------- 推荐存储：10G ---------- 推荐完成时间：1 周 ---------- 推荐网络配置：非实时 ---------- 其它配置	---网络--- 非实时 ---- 其他
2	-------- 计算能力： 8 路视频 清晰度 1080p ------- 存储：2T ------- 网络： 延迟<200ms, 上行带宽：xxMB 下行带宽：xxMb -------- 编解码能力： xxx 标准 -------	---计算单元--- 逻辑运算：3TOPS 矩阵运行：5TFLOPS 张量运算：-- ------存储------- 25T ------网络---------- 延迟<200ms, 上行带宽：xxMB 下行带宽：xxMb ------编解码---- 2 个解码引擎 4 个编码引擎 ------

[0085] 第一输出解释模块23用于对第一神经网络模型21得到预测的结果进行翻译，得到算力参数。

[0086] 示例性的，在确定算力业务请求的类型为业务需求类型时，将算力业务请求中的业务需求参数代入到第一神经网络模型中，得到第一预测结果，通过第一输出解释机制对第一预测结果进行解释，得到算力参数。

[0087] 第二神经网络模型与算力业务请求的类型为抽象需求类型相对应。即在算力业务请求的类型为抽象需求类型时，当前的神经网络模型为第二神经网络模型。

[0088] 第二神经网络模型用于根据抽象需求参数预测算力参数。

[0089] 可选的，第二神经网络模型可以为第二神经网络专家系统中的神经网络模型。

[0090] 如图3所示，第二神经网络专家系统30包括第二神经网络模型31、第二规则库32以及第二输出解释模块33。

[0091] 第二神经网络专家系统30的输入为算力业务请求中的抽象需求参数，第二神经网络专家系统30的输出为算力参数。

[0092] 第二规则库32用于保存抽象需求参数与算力参数对应的专家知识，并对第二神经网络模型31进行训练。

[0093] 可选的，第二规则库32可以保存从实际经验中获取的算力参数与业务需求参数对应的专家知识。专家知识可以是来自实际运行的大量实例集，也可以是从领域中的专家或文献资料中获取的大量规则，或者是二者的混合。

[0094] 示例性的，如表4所示，第二规则库32包括抽象需求参数为训练一个图像识别的网络模型、训练数据8G以及模型不确定，对应的算力参数为逻辑运算为10TOPS、矩阵运行为12TFLOPS、张量运算为8TFLOPS、存储为10G、网络为非实时；抽象需求参数为云增强现实(augmented reality，AR)/虚拟现实(virtual reality，VR)游戏的部署，游戏全部内容200G，对应的算力参数为，对应的算力参数为逻辑运算为3TOPS、矩阵运行为5TFLOPS、存储为25T、网络为延迟<200ms、上行带宽为xxMB、下行带宽为xxMb、编解码为2个解码引擎和4个编码引擎。

[0095] 表4

[0096]

序号	抽象需求参数	算力参数
1	训练一个图像识别的网络模型 训练数据 8G，模型不确定	计算单元 逻辑运算：10TOPS 矩阵运行：12TFLOPS 张量运算：8TFLOPS ------存储------- 10G ---网络--- 非实时 ---- 其他
2	云 AR/VR 游戏的部	---计算单元---

[0097]

	署，游戏全部内容200G	逻辑运算：3TOPS 矩阵运行：5TFLOPS 张量运算：-- ------存储------- 25T ------网络----------- 延迟<200ms, 上行带宽：xxMB 下行带宽：xxMb ------编解码---- 2个解码引擎 4个编码引擎 ------

[0098] 第二输出解释模块33用于将第二神经网络模型31得到预测的结果进行翻译，得到算力参数。

[0099] 示例性的，在确定算力业务请求的类型为抽象需求类型时，将算力业务请求中的抽象需求参数代入到第二神经网络模型中，得到第二预测结果，通过第二输出解释机制对第二预测结果进行解释，得到算力参数。

[0100] 可选的，结合图1，如图4所示，在S103之后，该算力参数的确定方法还包括：

[0101] S104、确定装置确定算力参数与目标算力参数的差值小于或等于预设阈值。

[0102] 目标算力参数为满足算力业务请求的算力参数。

[0103] 可选的，在确定装置确定出算力参数之后，可以将算力参数与目标算力参数进行对比，在算力参数与目标算力参数的差值小于或等于预设阈值时，确定算力参数满足用户的算力业务请求。

[0104] S105、基于算力参数和算力业务请求，优化当前的神经网络模型。

[0105] 一种可实现的方式，基于算力参数和算力业务请求中的业务需求参数，优化当前的神经网络模型。

[0106] 示例性的，在确定算力业务请求的类型为业务需求类型时，基于算力业务请求中的业务需求参数与第一神经网络模型，确定算力参数，且算力参数与目标算力参数的差值小于或等于预设阈值的时，根据确定出的算力参数以及算力业务请求中的业务需求参数，更新第一规则库。

[0107] 利用更新后的第一规则库，训练第一神经网络模型，得到训练后的第一神经网络模型。

[0108] 例如，如表5所示，更新后的第一规则库包括用户的抽象需求参数为训练一个图像

识别的网络模型训练数据8G，模型不确定，业务需求参数为推荐经典AlexNet模型配置、80w节点、6096w参数、卷积层参数3.8%、全连接层参数96.2%、推荐存储为10G、推荐完成时间为1周、推荐网络配置为非实时，对应的算力参数为逻辑运算为10TOPS、矩阵运行为12TFLOPS、张量运算为8TFLOPS、存储为10G、网络为非实时，且算力参数满足用户的需求。

[0109] 表5

[0110]

抽象需求参数	业务需求参数	算力参数	用户满意度
训练一个图像识别的网络模型 训练数据 8G， 模型不确定	---------- 计算能力: 推荐经典 AlexNet 模型配置: 80w 节点， 6096w 参数， 卷积层参数 3.8% 全连接层参数 96.2% ---------- 推荐存储：10G ---------- 推荐完成时间：1周 ---------- 推荐网络配置：非实时 ---------- 其它配置	-----计算单元--- 逻辑运算：10TOPS 矩阵运行：12TFLOPS 张量运算：8TFLOPS ------存储------- 10G ---网络--- 非实时 ---- 其他	Good

[0111] 在下一次确定装置基于算力业务请求中的业务需求参数与当前的神经网络模型，确定算力参数时，训练后的第一神经网络模型即为当前的神经网络模型。

[0112] 另一种可实现的方式，基于算力参数和算力业务请求中的抽象需求参数，优化当前的神经网络模型。

[0113] 示例性的，在确定算力业务请求的类型为抽象需求类型时，基于算力业务请求中的抽象需求参数与第二神经网络模型，确定算力参数。

[0114] 算力参数与目标算力参数的差值小于或等于预设阈值。则根据确定出的算力参数以及算力业务请求中的抽象需求参数，更新第二规则库。

[0115] 利用更新后的第二规则库，训练第二神经网络模型，得到训练后的第二神经网络

模型。

[0116] 在下一次确定装置基于算力业务请求中的抽象需求参数与当前的神经网络模型，确定算力参数时，训练后的第二神经网络模型即为当前的神经网络模型。

[0117] 本申请提供的算力参数的确定方法，通过获取用户的算力业务请求，基于算力业务请求与当前的神经网络模型，确定算力参数。相比于现有技术中人工确定业务请求对应的算力参数的方法，需要用户具备一定的专业能力，而且效率较低，本申请提供的算力参数的确定方法是基于神经网络专家系统，确定算力参数，从而不需要用户具备一定的专业能力，进而能够提高确定算力参数的效率。

[0118] 上述主要从方法的角度对本申请实施例提供的方案进行了介绍。为了实现上述功能，其包含了执行各个功能相应的硬件结构和/或软件模块。本领域技术人员应该很容易意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，本申请能够以硬件或硬件和计算机软件的结合形式来实现。某个功能究竟以硬件还是计算机软件驱动硬件的方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请的范围。

[0119] 如图5所示，本申请实施例提供一种算力参数的确定设备500。该算力参数的确定设备可以包括至少一个处理器501，通信线路502，存储器503，通信接口504。

[0120] 具体的，处理器501，用于执行存储器503中存储的计算机执行指令，从而实现终端的步骤或动作。

[0121] 处理器501可以是一个芯片。例如，可以是现场可编程门阵列(field programmable gate array，FPGA)，可以是专用集成芯片(application spe cific integrated circuit，ASIC)，还可以是系统芯片(system on chip，So C)，还可以是中央处理器(central processor unit，CPU)，还可以是网络处理器(network processor，NP)，还可以是数字信号处理电路(digital si gnal processor，DSP)，还可以是微控制器(micro controller unit，MCU)，还可以是可编程控制器(programmable logic device，PLD)或其他集成芯片。

[0122] 通信线路502，用于在上述处理器501与存储器503之间传输信息。

[0123] 存储器503，用于存储执行计算机执行指令，并由处理器501来控制执行。

[0124] 存储器503可以是独立存在，通过通信线路502与处理器相连接。存储器503可以是易失性存储器或非易失性存储器，或可包括易失性和非易失性存储器两者。其中，非易失性存储器可以是只读存储器(read-only memory，ROM)、可编程只读存储器(programmable ROM，PROM)、可擦除可编程只读存储器(erasable PROM，EPROM)、电可擦除可编程只读存储器(e lectrically EPROM，EEPROM)或闪存。易失性存储器可以是随机存取存储器(random access memory，RAM)，其用作外部高速缓存。通过示例性但不是限制性说明，许多形式的RAM可用，例如静态随机存取存储器(st atic RAM，SRAM)、动态随机存取存储器(dynamic RAM，DRAM)、同步动态随机存取存储器(synchronous DRAM，SDRAM)、双倍数据速率同步动态随机存取存储器(double data rate SDRAM，DDR SDRAM)、增强型同步动态随机存取存储器(enhanced SDRAM，ESDRAM)。应注意，本文描述的系统和设备的存储器旨在包括但不限于这些和任意其它适合类型的存储器。

[0125] 通信接口504，用于与其他设备或通信网络通信。其中，通信网络可以是以太网，无

线接入网(radio access network,RAN),或无线局域网(w ireless local area networks,WLAN)等。

[0126] 需要指出的是,图5中示出的结构并不构成对该算力参数的确定设备的限定,除图5所示部件之外,该算力参数的确定设备可以包括比图示更多或更少的部件,或者组合某些部件,或者不同的部件布置。

[0127] 如图6所示,本申请实施例提供一种算力参数的确定装置60。该算力参数的确定装置可以包括获取单元61、确定单元62。

[0128] 获取单元61,用于获取用户的算力业务请求。例如,结合图1,获取单元61可以用于执行S101。

[0129] 确定单元62,用于确定获取单元61获取的算力业务请求的类型。例如,结合图1,确定单元62可以用于执行S102。

[0130] 确定单元62,还用于基于获取单元61获取的算力业务请求与当前的神经网络模型,确定算力参数。例如,结合图1,确定单元62可以用于执行S103。

[0131] 应理解,在本申请的各种实施例中,上述各过程的序号的大小并不意味着执行顺序的先后,各过程的执行顺序应以其功能和内在逻辑确定,而不应对本申请实施例的实施过程构成任何限定。

[0132] 在实际实现时,获取单元61以及确定单元62,可以由图5所示的处理器501调用存储器503中的程序代码来实现。其具体的执行过程可参考图1所示的算力参数的确定方法部分的描述,这里不再赘述。

[0133] 本申请另一实施例还提供一种计算机可读存储介质,该计算机可读存储介质中存储有计算机指令,当计算机指令在算力参数的确定装置上运行时,使得在算力参数的确定装置执行上述方法实施例所示的方法流程中在算力参数的确定装置执行的各个步骤。

[0134] 在本申请另一实施例中,还提供一种计算机程序产品,该计算机程序产品包括指令,当指令在算力参数的确定装置上运行时,使得在算力参数的确定装置执行上述方法实施例所示的方法流程中在算力参数的确定装置执行的各个步骤。

[0135] 本领域普通技术人员可以意识到,结合本文中所公开的实施例描述的各示例的单元及算法步骤,能够以电子硬件、或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行,取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能,但是这种实现不应认为超出本申请的范围。

[0136] 所属领域的技术人员可以清楚地了解到,为描述的方便和简洁,上述描述的系统、设备和单元的具体工作过程,可以参考前述方法实施例中的对应过程,在此不再赘述。

[0137] 在本申请所提供的几个实施例中,应该理解到,所揭露的系统、设备和方法,可以通过其它的方式实现。例如,以上所描述的设备实施例仅仅是示意性的,例如,单元的划分,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,例如多个单元或组件可以结合或者可以集成到另一个系统,或一些特征可以忽略,或不执行。另一点,所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口,设备或单元的间接耦合或通信连接,可以是电性,机械或其它的形式。

[0138] 作为分离部件说明的单元可以是或者也可以不是物理上分开的,作为单元显示的

部件可以是或者也可以不是物理单元,即可以位于一个地方,或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0139] 另外,在本申请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。

[0140] 以上,仅为本申请的具体实施方式,但本申请的保护范围并不局限于此,任何熟悉本技术领域的技术人员在本申请揭露的技术范围内,可轻易想到的变化或替换,都应涵盖在本申请的保护范围之内。因此,本申请的保护范围应以权利要求的保护范围为准。

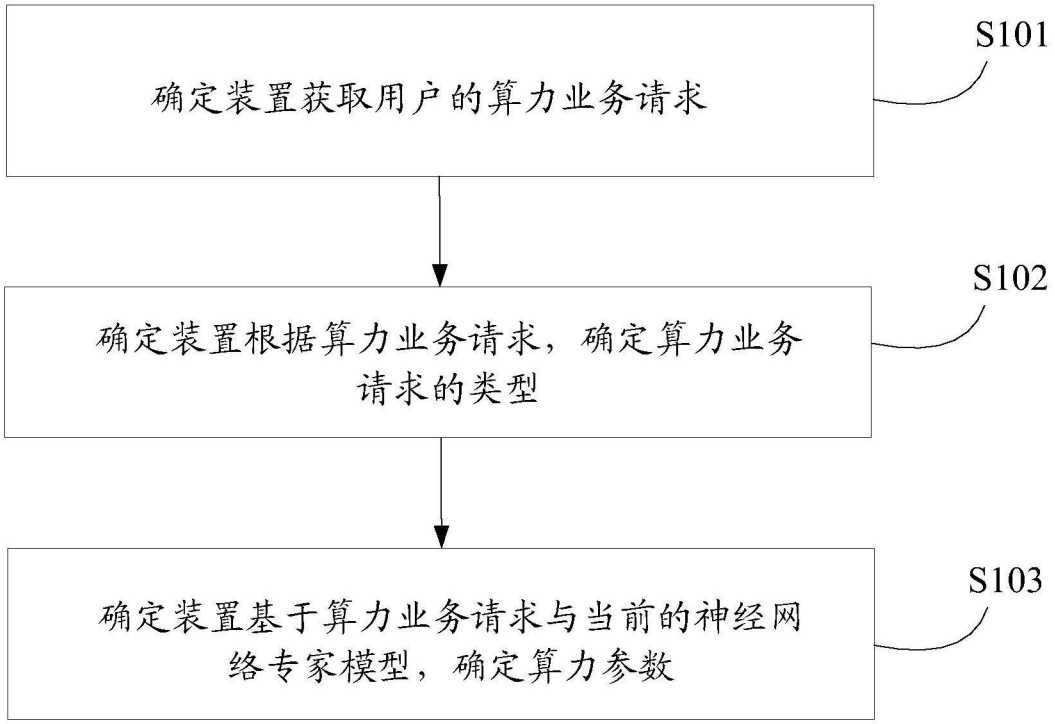

确定装置获取用户的算力业务请求
S101
确定装置根据算力业务请求，确定算力业务请求的类型
S102
确定装置基于算力业务请求与当前的神经网络专家模型，确定算力参数
S103

图1

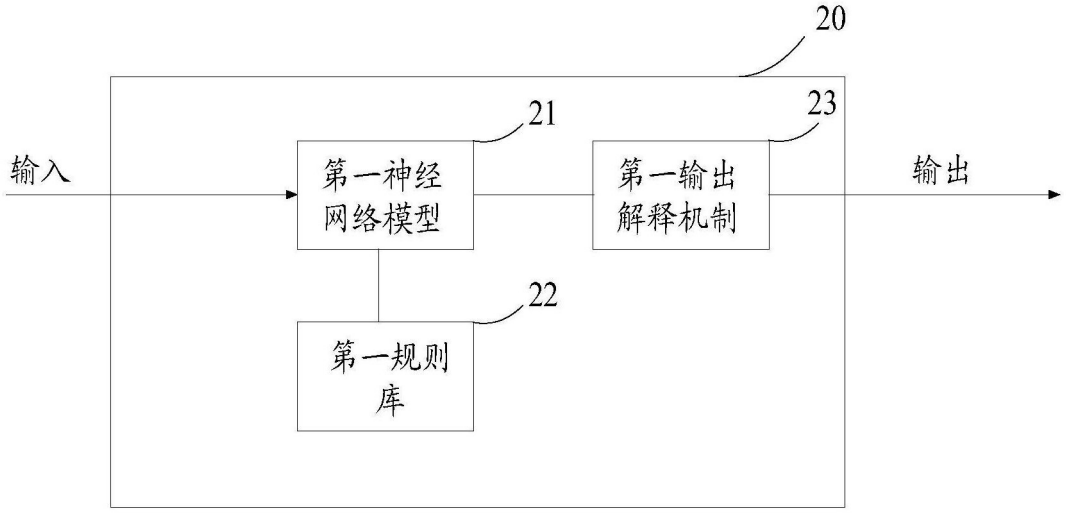

20
21
23
输入
第一神经网络模型
第一输出解释机制
输出
22
第一规则库

图2

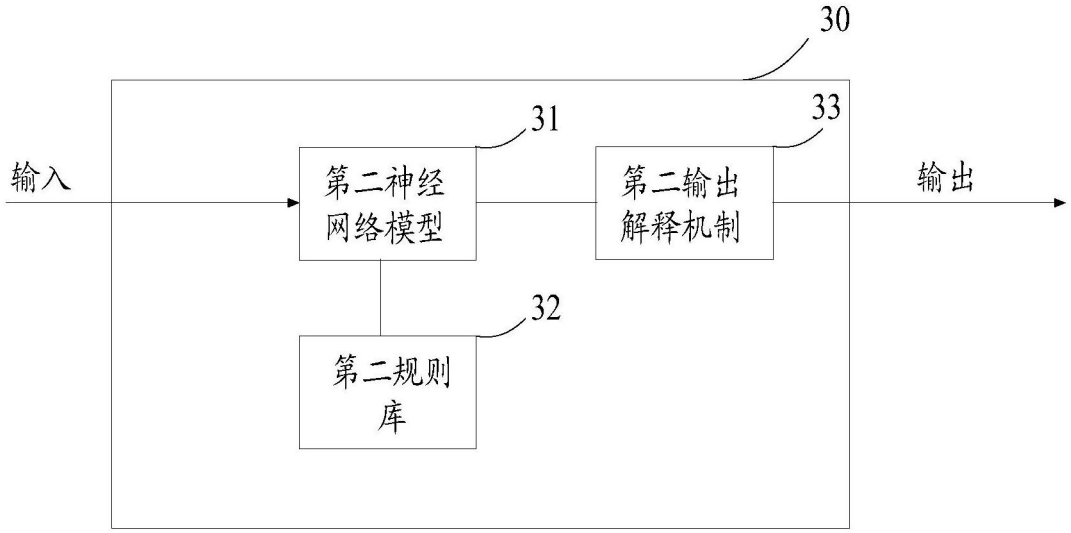


图3

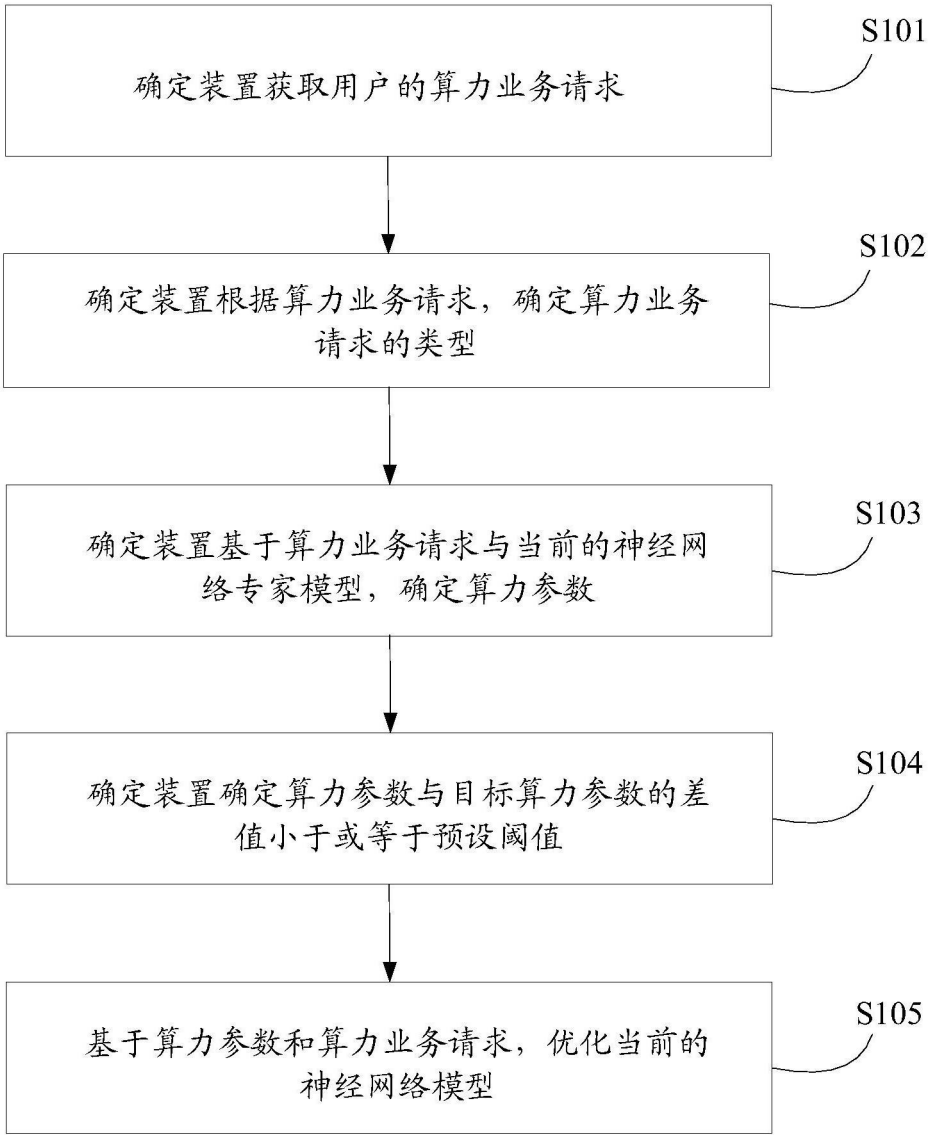

确定装置获取用户的算力业务请求
S101
确定装置根据算力业务请求，确定算力业务请求的类型
S102
确定装置基于算力业务请求与当前的神经网络专家模型，确定算力参数
S103
确定装置确定算力参数与目标算力参数的差值小于或等于预设阈值
S104
基于算力参数和算力业务请求，优化当前的神经网络模型
S105

图4

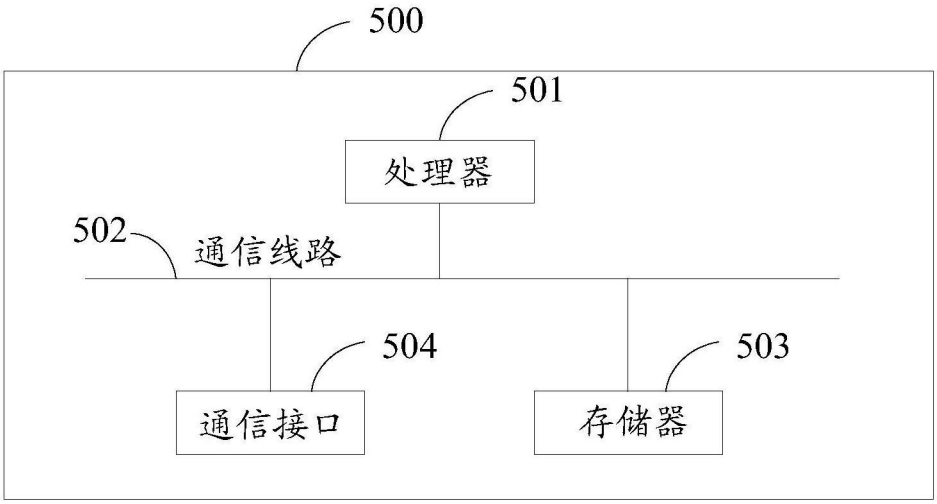


图5

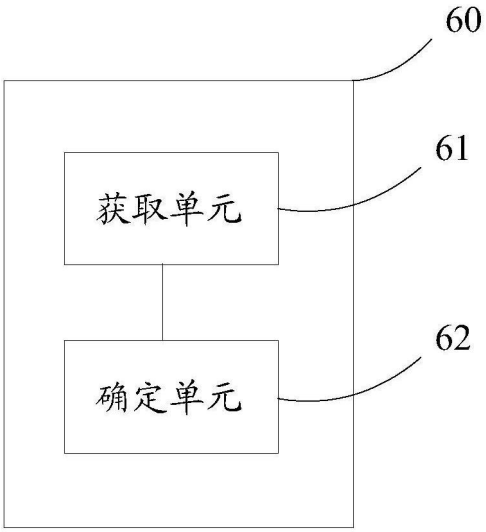


图6