(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 117311998 B
(45)授权公告日 2024.03.05

(21)申请号 202311618105.9

(22)申请日 2023.11.30

(65)同一申请的已公布的文献号
申请公布号 CN 117311998 A

(43)申请公布日 2023.12.29

(73)专利权人 卓世未来(天津)科技有限公司
地址 301726 天津市武清区武清开发区福源道北侧创业总部基地C02号楼313室33号

(72)发明人 屠静 王亚 赵策 潘亮亮 李伟伟 苏岳 孙岩 颉彬 万晶晶 周勤民

(74)专利代理机构 北京市广友专利事务所有限责任公司 11237
专利代理师 邓琳 张仲波

(51)Int.Cl.
G06F 9/50(2006.01)

(56)对比文件
CN 104579747 A,2015.04.29
CN 113220457 A,2021.08.06
CN 113723443 A,2021.11.30
CN 114897155 A,2022.08.12
CN 116204321 A,2023.06.02
CN 116450312 A,2023.07.18
CN 116801219 A,2023.09.22
US 2019392307 A1,2019.12.26
WO 2023123828 A1,2023.07.06

审查员 么旭君

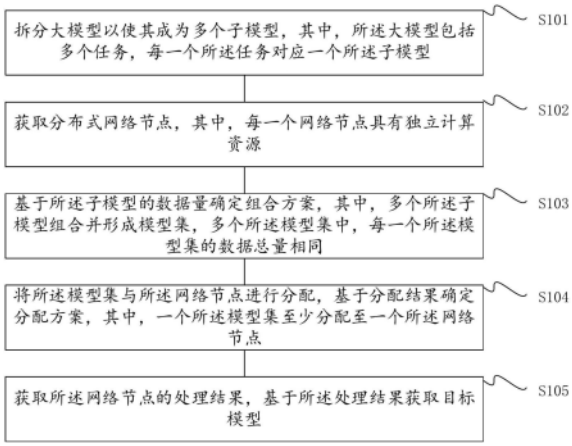
权利要求书2页 说明书9页 附图1页

(54)发明名称

一种大模型部署方法及系统

(57)摘要

本发明提供一种大模型部署方法及系统,属于人工智能技术领域。所述方法包括:拆分大模型以使其成为多个子模型,其中,所述大模型包括多个任务,每一个所述任务对应一个所述子模型;获取分布式网络节点,其中,每一个所述网络节点具有独立计算资源;基于所述子模型的数据量确定组合方案,其中,多个所述子模型组合并形成模型集,多个所述模型集中,每一个所述模型集的数据总量相同;将所述模型集与所述网络节点进行分配,基于分配结果确定分配方案,其中,一个所述模型集至少分配至一个所述网络节点;获取所述网络节点的处理结果,基于所述处理结果获取目标模型。本发明将大模型拆分后按条件分配至分布式网络节点,能有效提高计算效率和吞吐量。



CN 117311998 B

1.一种大模型部署方法,其特征在于,包括以下步骤:

拆分大模型以使其成为多个子模型,其中,所述大模型包括多个任务,每一个所述任务对应一个所述子模型;

获取分布式网络节点,其中,每一个所述网络节点具有独立计算资源;

基于所述子模型的数据量确定组合方案,其中,多个所述子模型组合并形成模型集,多个所述模型集中,每一个所述模型集的数据总量相同;

将所述模型集与所述网络节点进行分配,基于分配结果确定分配方案,其中,一个所述模型集至少分配至一个所述网络节点;

获取所述网络节点的处理结果,基于所述处理结果获取目标模型;

其中,所述拆分大模型以使其成为多个子模型,其中,所述大模型包括多个任务,每一个所述任务对应一个所述子模型,包括:

获取所述大模型中所有任务的对应输入、输出和目标;

基于每个所述任务中,所述对应输入、所述输出和所述目标的关联性,确定所述任务的权重;

基于所述权重对所述大模型进行拆分,以使其成为多个子任务;

其中,所述基于所述子模型的数据量确定组合方案,其中,多个所述子模型组合并形成模型集,多个所述模型集中,每一个所述模型集的数据总量相同,包括:

获取每一个所述子模型已知的数据量,并将多个所述子模型按数据量的数值进行梯度排序;

根据所述梯度排序的顺序及分布情况,确定未知的所述子模型的数据量;

根据预设条件合并多个所述子模型为一个所述模型集,每一个所述模型集的数据总量相同;

其中,所述根据所述梯度排序的顺序及分布情况,确定未知的所述子模型的数据量,包括:

基于所述梯度排序确定多个所述子模型的数据量曲线;

基于所述数据量曲线推测未知的所述子模型的数据量,具体为确定所述未知的子模型的数据量在所述数据量曲线中沿正态分布的位置。

2.根据权利要求1所述的大模型部署方法,其特征在于,所述根据预设条件合并多个所述子模型为一个所述模型集,每一个所述模型集的数据总量相同,包括:

若所述子模型的数量大于所述网络节点数量,则基于所述网络节点数量确定所述模型集的数量,其中,所述预设条件为所述模型集的数量等于所述网络节点数量。

3.根据权利要求2所述的大模型部署方法,其特征在于,所述若所述子模型的数量大于所述网络节点数量,则基于所述网络节点数量确定所述模型集的数量,其中,所述预设条件为所述模型集的数量等于所述网络节点数量,包括:

根据每一个所述子模型的数据量,合并多个所述子模型,以使所述子模型集的数量等于所述网络节点数量。

4.一种大模型部署系统,其特征在于,所述系统包括:

第一拆分模块,所述第一拆分模块用于拆分大模型以使其成为多个子模型,其中,所述大模型包括多个任务,每一个所述任务对应一个所述子模型;

第一获取模块,所述第一获取模块用于获取分布式网络节点,其中,每一个网络节点具有独立计算资源;

第一确定模块,所述第一确定模块用于基于所述子模型的数据量确定组合方案,其中,多个所述子模型组合并形成模型集,多个所述模型集中,每一个所述模型集的数据总量相同;

第一分配模块,所述第一分配模块用于将所述模型集与所述网络节点进行分配,基于分配结果确定分配方案,其中,一个所述模型集至少分配至一个所述网络节点;

第二获取模块,所述第二获取模块用于获取所述网络节点的处理结果,基于所述处理结果获取目标模型;

其中,所述系统还包括:

第四获取模块,所述第四获取模块用于获取所述大模型中所有任务的对应输入、输出和目标;

第四确定模块,所述第四确定模块用于基于每个所述任务中,所述对应输入、所述输出和所述目标的关联性,确定所述任务的权重;

第二拆分模块,所述第二拆分模块用于基于所述权重对所述大模型进行拆分,以使其成为多个子任务;

其中,所述系统还包括:

第三获取模块,所述第三获取模块用于获取每一个所述子模型已知的数据量,并将多个所述子模型按数据量的数值进行梯度排序;

第二确定模块,所述第二确定模块用于根据所述梯度排序的顺序及分布情况,确定未知的所述子模型的数据量;

第一合并模块,所述第一合并模块用于根据预设条件合并多个所述子模型为一个所述模型集,每一个所述模型集的数据总量相同;

其中,所述系统还包括:

第五确定模块,所述第五确定模块用于基于所述梯度排序确定多个所述子模型的数据量曲线;

预测模块,所述预测模块用于基于所述数据量曲线推测未知的所述子模型的数据量,具体为确定所述未知的子模型的数据量在所述数据量曲线中沿正态分布的位置。

5.一种电子设备,其特征在于,包括处理器、通信接口、存储器和通信总线,其中,所述处理器,所述通信接口,所述存储器通过所述通信总线完成相互间的通信;

所述存储器,用于存放计算机程序;

所述处理器,用于执行所述存储器上所存放的程序时,实现如权利要求1-3任一项所述的方法。

6.一种计算机可读存储介质,其上存储有计算机程序,该程序被处理器执行时实现如权利要求1-3任一项所述的方法。

一种大模型部署方法及系统

技术领域

[0001] 本发明涉及人工智能技术领域,尤其涉及一种大模型部署方法及系统。

背景技术

[0002] 大模型部署是指将训练好的大型模型应用到实际生产环境中的过程。模型通常需要大量的存储和计算资源,因此在部署之前需要对模型进行压缩和优化。常用的技术包括参数剪枝、量化、低秩分解等,以减少模型的大小和计算量,同时保持模型的性能。为了进一步提高大模型的性能,可以采用硬件加速和优化技术。例如,使用图形处理器(GPU)或专用的神经网络处理器(如Google的TPU)来加速模型的计算速度。此外,对硬件和软件进行深度优化,如使用高效的矩阵运算库(如cuDNN)和模型推理引擎(如TensorRT),可以提高模型的运行效率。

[0003] 大模型的训练和推理往往需要巨大的计算能力,单个计算节点难以满足需求。因此,采用分布式计算技术可以将计算任务分布到多个计算节点上并行处理,提高计算效率和吞吐量。相关技术中的大模型部署方法,直接将大模型分配至分布式网络节点,计算效率和吞吐量较低。

发明内容

[0004] 针对上述问题,本发明的目的在于提供一种大模型部署方法及系统,以解决现有技术中的大模型部署方法计算效率和吞吐量较低的问题。

[0005] 为解决上述技术问题,本发明提供如下技术方案:

[0006] 第一方面,提供一种大模型部署方法,包括以下步骤:

[0007] 拆分大模型以使其成为多个子模型,其中,所述大模型包括多个任务,每一个所述任务对应一个所述子模型;

[0008] 获取分布式网络节点,其中,每一个所述网络节点具有独立计算资源;

[0009] 基于所述子模型的数据量确定组合方案,其中,多个所述子模型组合并形成模型集,多个所述模型集中,每一个所述模型集的数据总量相同;

[0010] 将所述模型集与所述网络节点进行分配,基于分配结果确定分配方案,其中,一个所述模型集至少分配至一个所述网络节点;

[0011] 获取所述网络节点的处理结果,基于所述处理结果获取目标模型。

[0012] 结合第一方面,在一些实施方式中,所述基于所述子模型的数据量确定组合方案,其中,多个所述子模型组合并形成模型集,多个所述模型集中,每一个所述模型集的数据总量相同,包括:

[0013] 获取每一个所述子模型已知的数据量,并将多个所述子模型按数据量的数值进行梯度排序;

[0014] 根据所述梯度排序的顺序及分布情况,确定未知的所述子模型的数据量;

[0015] 根据预设条件合并多个所述子模型为一个所述模型集,每一个所述模型集的数据

总量相同。

[0016] 结合第一方面,在一些实施方式中,所述根据预设条件合并多个所述子模型为一个所述模型集,每一个所述模型集的数据总量相同,包括:

[0017] 若所述子模型的数量大于所述网络节点数量,则基于所述网络节点数量确定所述模型集的数量,其中,所述预设条件为所述模型集的数量等于所述网络节点数量。

[0018] 结合第一方面,在一些实施方式中,所述若所述子模型的数量大于所述网络节点数量,则基于所述网络节点数量确定所述模型集的数量,其中,所述预设条件为所述模型集的数量等于所述网络节点数量,包括:

[0019] 根据每一个所述子模型的数据量,合并多个所述子模型,以使所述子模型集的数量等于所述网络节点数量。

[0020] 结合第一方面,在一些实施方式中,所述拆分大模型以使其成为多个子模型,其中,所述大模型包括多个任务,每一个所述任务对应一个所述子模型,包括:

[0021] 获取所述大模型中所有任务的对应输入、输出和目标;

[0022] 基于每个所述任务中,所述对应输入、所述输出和所述目标的关联性,确定所述任务的权重;

[0023] 基于所述权重对所述大模型进行拆分,以使其成为多个所述子任务。

[0024] 结合第一方面,在一些实施方式中,所述根据所述梯度排序的顺序及分布情况,确定未知的所述子模型的数据量,包括:

[0025] 基于所述梯度排序确定多个所述子模型的数据量曲线;

[0026] 基于所述数据量曲线推测未知的所述子模型的数据量,具体为确定所述未知的子模型的数据量在所述数据量曲线中沿正态分布的位置。

[0027] 第二方面,提供一种大模型部署系统,所述系统包括:

[0028] 第一拆分模块,所述第一拆分模块用于拆分大模型以使其成为多个子模型,其中,所述大模型包括多个任务,每一个所述任务对应一个所述子模型;

[0029] 第一获取模块,所述第一获取模块用于获取分布式网络节点,其中,每一个网络节点具有独立计算资源;

[0030] 第一确定模块,所述第一确定模块用于基于所述子模型的数据量确定组合方案,其中,多个所述子模型组合并形成模型集,多个所述模型集中,每一个所述模型集的数据总量相同;

[0031] 第一分配模块,所述第一分配模块用于将所述模型集与所述网络节点进行分配,基于分配结果确定分配方案,其中,一个所述模型集至少分配至一个所述网络节点;

[0032] 第二获取模块,所述第二获取模块用于获取所述网络节点的处理结果,基于所述处理结果获取目标模型。

[0033] 结合第二方面,在一些实施方式中,所述系统还包括:

[0034] 第三获取模块,所述第三获取模块用于获取每一个所述子模型已知的数据量,并将多个所述子模型按数据量的数值进行梯度排序;

[0035] 第二确定模块,所述第二确定模块用于根据所述梯度排序的顺序及分布情况,确定未知的所述子模型的数据量;

[0036] 第一合并模块,所述第一合并模块用于根据预设条件合并多个所述子模型为一个

所述模型集,每一个所述模型集的数据总量相同。

[0037] 结合第二方面,在一些实施方式中,所述系统还包括:包括:

[0038] 第三确定模块,所述第三确定模块用于若所述子模型的数量大于所述网络节点数量,则基于所述网络节点数量确定所述模型集的数量,其中,所述预设条件为所述模型集的数量等于所述网络节点数量。

[0039] 结合第二方面,在一些实施方式中,所述系统还包括:

[0040] 第二合并模块,所述第二合并模块用于根据每一个所述子模型的数据量,合并多个所述子模型,以使所述子模型集的数量等于所述网络节点数量。

[0041] 结合第二方面,在一些实施方式中,所述系统还包括:

[0042] 第四获取模块,所述第四获取模块用于获取所述大模型中所有任务的对应输入、输出和目标;

[0043] 第四确定模块,所述第四确定模块用于基于每个所述任务中,所述对应输入、所述输出和所述目标的关联性,确定所述任务的权重;

[0044] 第二拆分模块,所述第二拆分模块用于基于所述权重对所述大模型进行拆分,以使其成为多个所述子任务。

[0045] 结合第二方面,在一些实施方式中,所述系统还包括:

[0046] 第五确定模块,所述第五确定模块用于基于所述梯度排序确定多个所述子模型的数据量曲线;

[0047] 预测模块,所述预测模块用于基于所述数据量曲线推测未知的所述子模型的数据量,具体为确定所述未知的子模型的数据量在所述数据量曲线中沿正态分布的位置。

[0048] 第三方面,提供一种电子设备,包括处理器、通信接口、存储器和通信总线,其中,处理器,通信接口,存储器通过通信总线完成相互间的通信;

[0049] 存储器,用于存放计算机程序;

[0050] 处理器,用于执行存储器上所存放的程序时,实现本发明实施例第一方面提出方法步骤。

[0051] 第四方面,提供一种计算机可读存储介质,其上存储有计算机程序,该程序被处理器执行时实现如本发明实施例第一方面提出方法。

[0052] 与现有技术相比,本发明提供的技术方案具有以下有益效果:

[0053] 本发明提出的一种大模型部署方法,首先拆分大模型以使其成为多个子模型,其中,所述大模型包括多个任务,每一个所述任务对应一个所述子模型;获取分布式网络节点,其中,每一个网络节点具有独立计算资源,然后,基于所述子模型的数据量确定组合方案,其中,多个所述子模型组合并形成模型集,多个所述模型集中,每一个所述模型集的数据总量相同,然后,将所述模型集与所述网络节点进行分配,基于分配结果确定分配方案,其中,一个所述模型集至少分配至一个所述网络节点,最后,获取所述网络节点的处理结果,基于所述处理结果获取目标模型。本发明提出的大模型部署方法,将大模型拆分后并按条件分配至分布式网络节点,能有效提高计算效率和吞吐量。

附图说明

[0054] 为了更清楚地说明本发明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发明的一些实施例,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。

[0055] 图1是本发明实施例提供的一种大模型部署方法的流程示意图。

具体实施方式

[0056] 为使本发明实施例的目的、技术方案和优点更加清楚,下面将结合本发明实施例的附图,对本发明实施例的技术方案进行清楚、完整地描述。显然,所描述的实施例是本发明的一部分实施例,而不是全部的实施例。基于所描述的本发明的实施例,本领域普通技术人员在无需创造性劳动的前提下所获得的所有其他实施例,都属于本发明保护的范围。

[0057] 本发明实施例提出了一种大模型部署方法,如图1所示,所述方法包括以下步骤:

[0058] S101:拆分大模型以使其成为多个子模型,其中,所述大模型包括多个任务,每一个所述任务对应一个所述子模型。

[0059] 可以理解的,将大模型拆分成多个子模型后,可以将这些子模型分布式地部署在不同的计算节点上进行计算。这样可以充分利用分布式计算的优势,提高计算效率和并行性。

[0060] 具体的,作为一种实施方式,步骤S101可以包括以下步骤:

[0061] S101-1:获取所述大模型中所有任务的对应输入、输出和目标。

[0062] 可以理解的,每个子模型将专注于解决单个任务,并根据其特定的输入、输出和目标进行训练和部署。

[0063] S101-2:基于每个所述任务中,所述对应输入、所述输出和所述目标的关联性,确定所述任务的权重。

[0064] 可以理解的,可以根据历史数据或者实验结果来分析任务的关联性和影响程度,进而确定任务的权重。例如,可以通过统计分析输入和输出之间的关系、目标的达成情况等来确定任务的权重。

[0065] S101-3:基于所述权重对所述大模型进行拆分,以使其成为多个所述子任务。

[0066] 需要说明的是,确定任务权重是一个复杂的问题,需要综合考虑多个因素。不同任务的权重可能会随着时间和环境的变化而变化,因此需要定期评估和更新任务的权重。

[0067] S102:获取分布式网络节点,其中,每一个网络节点具有独立计算资源。

[0068] 可以理解的,通过网络进行通信和协作。这种分布式计算模型可以提供高性能、高可扩展性和高可用性的计算能力。

[0069] S103:基于所述子模型的数据量确定组合方案,其中,多个所述子模型组合并形成模型集,多个所述模型集中,每一个所述模型集的数据总量相同。

[0070] 在一些实施方式中,步骤S103包括以下步骤:

[0071] S103-1:获取每一个所述子模型已知的数据量,并将多个所述子模型按数据量的数值进行梯度排序;

[0072] S103-2:根据所述梯度排序的顺序及分布情况,确定未知的所述子模型的数据量。

[0073] 具体的,作为一种实施方式,步骤S103-2包括以下步骤:

[0074] S103-2-1:基于所述梯度排序确定多个所述子模型的数据量曲线;

[0075] S103-2-2:基于所述数据量曲线推测未知的所述子模型的数据量,具体为确定所述未知的子模型的数据量在所述数据量曲线中沿正态分布的位置。

[0076] 可以理解的,要推测未知的子模型的数据量,可以使用曲线拟合技术来估计其在曲线中的位置。由于未知子模型的数据量沿正态分布,可以使用正态分布的参数来进行估计。

[0077] 具体的,收集已知子模型的数据量和相应的曲线数据。这些已知数据可以用来构建一个数据量曲线模型。然后,选择一个合适的正态分布拟合方法,可以使用最小二乘法或最大似然估计等方法。通过拟合已知子模型的数据量和相应的曲线数据,可以得到正态分布的参数,包括均值和标准差。最后,使用得到的正态分布参数来计算未知子模型的数据量在曲线中的位置。可以使用概率密度函数计算未知子模型的数据量在曲线上的概率密度,然后根据概率密度的大小来确定其在曲线中的位置。

[0078] S103-3:根据预设条件合并多个所述子模型为一个所述模型集,每一个所述模型集的数据总量相同。

[0079] 具体的,作为一种实施方式,步骤S103-3包括以下步骤:

[0080] S103-3-1:若所述子模型的数量大于所述网络节点数量,则基于所述网络节点数量确定所述模型集的数量,其中,所述预设条件为所述模型集的数量等于所述网络节点数量。

[0081] 可以理解的,根据每一个所述子模型的数据量,合并多个所述子模型,以使所述子模型集的数量等于所述网络节点数量。

[0082] S104:将所述模型集与所述网络节点进行分配,基于分配结果确定分配方案,其中,一个所述模型集至少分配至一个所述网络节点。

[0083] S105:获取所述网络节点的处理结果,基于所述处理结果获取目标模型。

[0084] 本发明提出的一种大模型部署方法,首先拆分大模型以使其成为多个子模型,其中,所述大模型包括多个任务,每一个所述任务对应一个所述子模型;获取分布式网络节点,其中,每一个网络节点具有独立计算资源,然后,基于所述子模型的数据量确定组合方案,其中,多个所述子模型组合并形成模型集,多个所述模型集中,每一个所述模型集的数据总量相同,然后,将所述模型集与所述网络节点进行分配,基于分配结果确定分配方案,其中,一个所述模型集至少分配至一个所述网络节点,最后,获取所述网络节点的处理结果,基于所述处理结果获取目标模型。本发明提出的大模型部署方法,将大模型拆分后并按条件分配至分布式网络节点,能有效提高计算效率和吞吐量。

[0085] 基于同一发明构思,本申请提出了一种大模型部署系统,所述系统包括:

[0086] 第一拆分模块,所述第一拆分模块用于拆分大模型以使其成为多个子模型,其中,所述大模型包括多个任务,每一个所述任务对应一个所述子模型;

[0087] 第一获取模块,所述第一获取模块用于获取分布式网络节点,其中,每一个网络节点具有独立计算资源;

[0088] 第一确定模块,所述第一确定模块用于基于所述子模型的数据量确定组合方案,其中,多个所述子模型组合并形成模型集,多个所述模型集中,每一个所述模型集的数据总量相同;

[0089] 第一分配模块,所述第一分配模块用于将所述模型集与所述网络节点进行分配,

基于分配结果确定分配方案,其中,一个所述模型集至少分配至一个所述网络节点;

[0090] 第二获取模块,所述第二获取模块用于获取所述网络节点的处理结果,基于所述处理结果获取目标模型。

[0091] 在一些实施方式中,所述系统还包括:

[0092] 第三获取模块,所述第三获取模块用于获取每一个所述子模型已知的数据量,并将多个所述子模型按数据量的数值进行梯度排序;

[0093] 第二确定模块,所述第二确定模块用于根据所述梯度排序的顺序及分布情况,确定未知的所述子模型的数据量;

[0094] 第一合并模块,所述第一合并模块用于根据预设条件合并多个所述子模型为一个所述模型集,每一个所述模型集的数据总量相同。

[0095] 在一些实施方式中,所述系统还包括:

[0096] 第三确定模块,所述第三确定模块用于若所述子模型的数量大于所述网络节点数量,则基于所述网络节点数量确定所述模型集的数量,其中,所述预设条件为所述模型集的数量等于所述网络节点数量。

[0097] 在一些实施方式中,所述系统还包括:

[0098] 第二合并模块,所述第二合并模块用于根据每一个所述子模型的数据量,合并多个所述子模型,以使所述子模型集的数量等于所述网络节点数量。

[0099] 在一些实施方式中,所述系统还包括:

[0100] 第四获取模块,所述第四获取模块用于获取所述大模型中所有任务的对应输入、输出和目标;

[0101] 第四确定模块,所述第四确定模块用于基于每个所述任务中,所述对应输入、所述输出和所述目标的关联性,确定所述任务的权重;

[0102] 第二拆分模块,所述第二拆分模块用于基于所述权重对所述大模型进行拆分,以使其成为多个所述子任务。

[0103] 在一些实施方式中,所述系统还包括:

[0104] 第五确定模块,所述第五确定模块用于基于所述梯度排序确定多个所述子模型的数据量曲线;

[0105] 预测模块,所述预测模块用于基于所述数据量曲线推测未知的所述子模型的数据量,具体为确定所述未知的子模型的数据量在所述数据量曲线中沿正态分布的位置。

[0106] 本发明提出的一种大模型部署系统,首先拆分大模型以使其成为多个子模型,其中,所述大模型包括多个任务,每一个所述任务对应一个所述子模型;获取分布式网络节点,其中,每一个网络节点具有独立计算资源,然后,基于所述子模型的数据量确定组合方案,其中,多个所述子模型组合并形成模型集,多个所述模型集中,每一个所述模型集的数据总量相同,然后,将所述模型集与所述网络节点进行分配,基于分配结果确定分配方案,其中,一个所述模型集至少分配至一个所述网络节点,最后,获取所述网络节点的处理结果,基于所述处理结果获取目标模型。本发明提出的大模型部署系统,将大模型拆分后并按条件分配至分布式网络节点,能有效提高计算效率和吞吐量。

[0107] 基于同一发明构思,本申请的实施例还提出了一种电子设备,电子设备包括:

[0108] 至少一个处理器;以及,与至少一个处理器通信连接的存储器;其中,存储器存储

有可被至少一个处理器执行的指令,指令被至少一个处理器执行,以使至少一个处理器能够执行本申请实施例的大模型部署方法。

[0109] 此外,为实现上述目的,本申请的实施例还提出了一种计算机可读存储介质,存储有计算机程序,计算机程序被处理器执行时实现本申请实施例的大模型部署方法。

[0110] 下面对电子设备的各个构成部件进行具体的介绍:

[0111] 其中,处理器是电子设备的控制中心,可以是一个处理器,也可以是多个处理元件的统称。例如,处理器是一个或多个中央处理器(central processing unit,CPU),也可以是特定集成电路(application specific integrated circuit,ASIC),或者是被配置成实施本发明实施例的一个或多个集成电路,例如:一个或多个微处理器(digital signal processor,DSP),或,一个或者多个现场可编程门阵列(field programmable gate array,FPGA)。

[0112] 可选地,处理器可以通过运行或执行存储在存储器内的软件程序,以及调用存储在存储器内的数据,执行电子设备的各种功能。

[0113] 其中,所述存储器用于存储执行本发明方案的软件程序,并由处理器来控制执行,具体实现方式可以参考上述方法实施例,此处不再赘述。

[0114] 可选地,存储器可以是只读存储器(read-only memory,ROM)或可存储静态信息和指令的其他类型的静态存储设备,随机存取存储器(random access memory,RAM)或者可存储信息和指令的其他类型的动态存储设备,也可以是电可擦可编程只读存储器(electrically erasable programmable read-only memory,EEPROM)、只读光盘(compact disc read-only memory,CD-ROM)或其他光盘存储、光碟存储(包括压缩光碟、激光碟、光碟、数字通用光碟、蓝光光碟等)、磁盘存储介质或者其他磁存储设备、或者能够用于携带或存储具有指令或数据结构形式的期望的程序代码并能够由计算机存取的任何其他介质,但不限于此。存储器可以和处理器集成在一起,也可以独立存在,并通过电子设备的接口电路与处理器耦合,本发明实施例对此不作具体限定。

[0115] 收发器,用于与网络设备通信,或者与终端设备通信。

[0116] 可选地,收发器可以包括接收器和发送器。其中,接收器用于实现接收功能,发送器用于实现发送功能。

[0117] 可选地,收发器可以和处理器集成在一起,也可以独立存在,并通过路由器的接口电路与处理器耦合,本发明实施例对此不作具体限定。

[0118] 此外,电子设备的技术效果可以参考上述方法实施例所述的数据传输方法的技术效果,此处不再赘述。

[0119] 应理解,在本发明实施例中的处理器可以是中央处理单元(central processing unit,CPU),该处理器还可以是其他通用处理器、数字信号处理器(digital signal processor,DSP)、专用集成电路(application specific integrated circuit,ASIC)、现成可编程门阵列(field programmable gate array,FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

[0120] 还应理解,本发明实施例中的存储器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(electrically EPROM,EEPROM)或闪存。易失性存储器可以是随机存取存储器(random access memory,RAM),其用作外部高速缓存。通过示例性但不是限制性说明,许多形式的随机存取存储器(random access memory,RAM)可用,例如静态随机存取存储器(static RAM,SRAM)、动态随机存取存储器(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)、同步连接动态随机存取存储器(synchlink DRAM,SLDRAM)和直接内存总线随机存取存储器(direct rambus RAM,DR RAM)。

[0121] 上述实施例,可以全部或部分地通过软件、硬件(如电路)、固件或其他任意组合来实现。当使用软件实现时,上述实施例可以全部或部分地以计算机程序产品的形式实现。所述计算机程序产品包括一个或多个计算机指令或计算机程序。在计算机上加载或执行所述计算机指令或计算机程序时,全部或部分地产生按照本发明实施例所述的流程或功能。所述计算机可以为通用计算机、专用计算机、计算机网络、或者其他可编程装置。所述计算机指令可以存储在计算机可读存储介质中,或者从一个计算机可读存储介质向另一个计算机可读存储介质传输,例如,所述计算机指令可以从一个网站站点、计算机、服务器或数据中心通过有线(例如红外、无线、微波等)方式向另一个网站站点、计算机、服务器或数据中心进行传输。所述计算机可读存储介质可以是计算机能够存取的任何可用介质或者是包含一个或多个可用介质集合的服务器、数据中心等数据存储设备。所述可用介质可以是磁性介质(例如,软盘、硬盘、磁带)、光介质(例如,DVD)、或者半导体介质。半导体介质可以是固态硬盘。

[0122] 应理解,本文中术语“和/或”,仅仅是一种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,其中A,B可以是单数或者复数。另外,本文中字符“/”,一般表示前后关联对象是一种“或”的关系,但也可能表示的是一种“和/或”的关系,具体可参考前后文进行理解。

[0123] 本发明中,“至少一个”是指一个或者多个,“多个”是指两个或两个以上。“以下至少一项(个)”或其类似表达,是指的这些项中的任意组合,包括单项(个)或复数项(个)的任意组合。例如,a,b,或c中的至少一项(个),可以表示:a, b, c, a-b, a-c, b-c, 或a-b-c,其中a,b,c可以是单个,也可以是多个。

[0124] 应理解,在本发明的各种实施例中,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,而不应对本发明实施例的实施过程构成任何限定。

[0125] 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发明的范围。

[0126] 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、装置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。

[0127] 在本发明所提供的几个实施例中,应该理解到,所揭露的系统、装置和方法,可以通过其它的方式实现。例如,以上所描述的装置实施例仅仅是示意性的,例如,所述单元的划分,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,装置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。

[0128] 所述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。

[0129] 另外,在本发明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。

[0130] 所述功能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时,可以存储在一个计算机可读取存储介质中。基于这样的理解,本发明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)执行本发明各个实施例所述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。

[0131] 以上所述仅为本发明的较佳实施例,并不用以限制本发明,凡在本发明的精神和原则之内,所作的任何修改、等同替换、改进等,均应包含在本发明的保护范围之内。

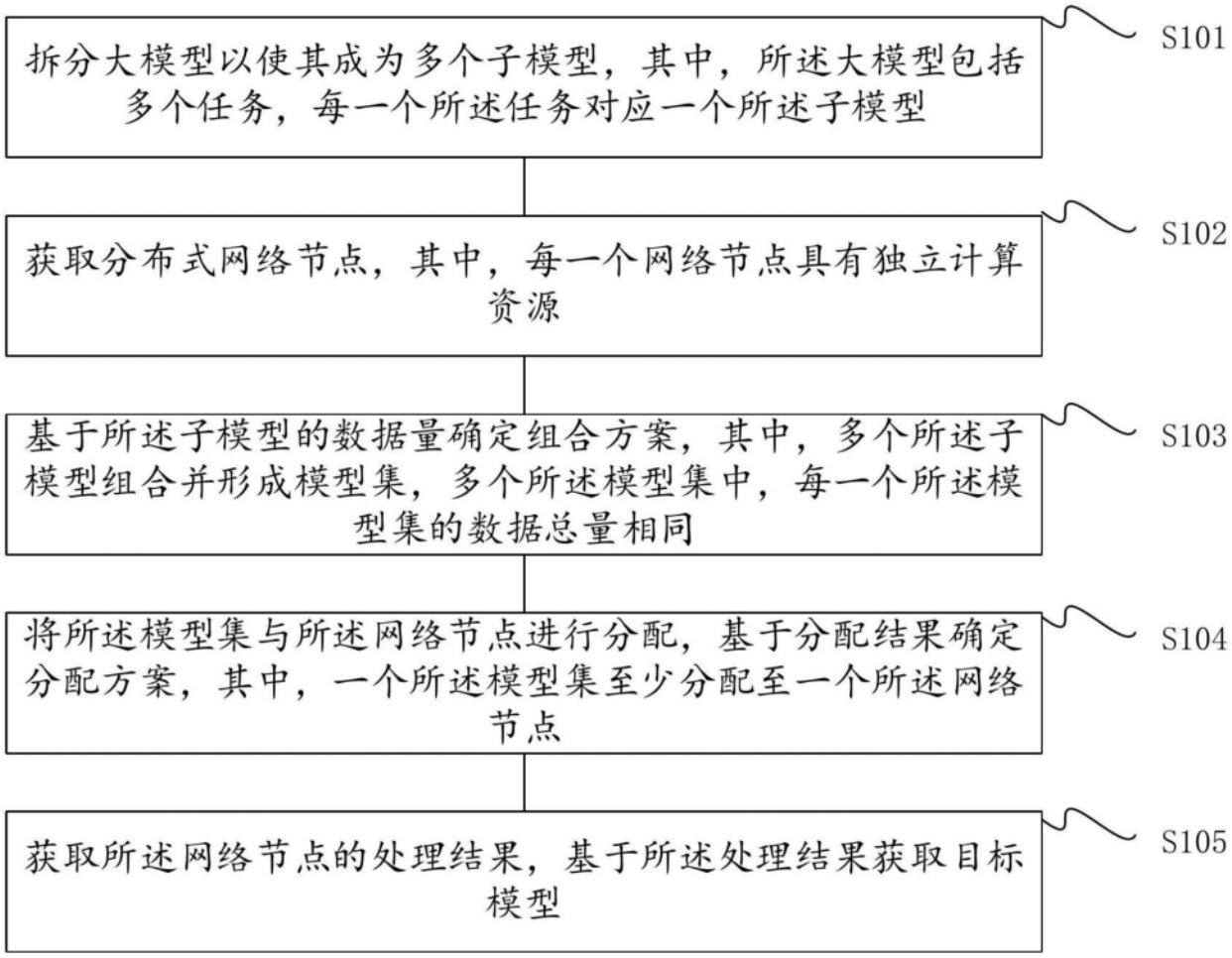

拆分大模型以使其成为多个子模型，其中，所述大模型包括多个任务，每一个所述任务对应一个所述子模型
S101
获取分布式网络节点，其中，每一个网络节点具有独立计算资源
S102
基于所述子模型的数据量确定组合方案，其中，多个所述子模型组合并形成模型集，多个所述模型集中，每一个所述模型集的数据总量相同
S103
将所述模型集与所述网络节点进行分配，基于分配结果确定分配方案，其中，一个所述模型集至少分配至一个所述网络节点
S104
获取所述网络节点的处理结果，基于所述处理结果获取目标模型
S105

图1