



(12) 发明专利

(10) 授权公告号 CN 113723552 B

(45) 授权公告日 2024. 11. 08

(21) 申请号 202111042840.0

G06N 3/088 (2023.01)

(22) 申请日 2021.09.07

G06F 9/50 (2006.01)

(65) 同一申请的已公布的文献号

申请公布号 CN 113723552 A

(56) 对比文件

CN 111814911 A, 2020.10.23

CN 112417358 A, 2021.02.26

(43) 申请公布日 2021.11.30

审查员 刘梦佳

(73) 专利权人 北京大学深圳研究生院

地址 518055 广东省深圳市南山区西丽镇

丽水路深圳大学城北大校区

(72) 发明人 李革 任俞睿 王耀威 白鑫贝

郭明月

(74) 专利代理机构 北京京万通知识产权代理有

限公司 11440

专利代理师 万学堂

(51) Int. Cl.

G06F 18/214 (2023.01)

权利要求书2页 说明书12页 附图2页

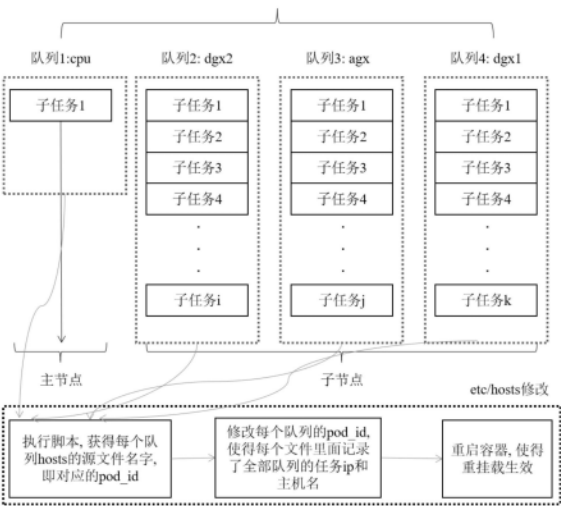
(54) 发明名称

大规模多机多卡预训练方法、系统、设备及服务器集群

(57) 摘要

本发明属于分布式训练技术领域,公开了一种大规模多机多卡预训练方法、系统、设备及服务器集群,在多个服务器上部署多机多卡,进行同构机型和异构混合机型的多机多卡并行;基于slurm框架进行大规模多机多卡训练及评测,以无监督特征学习BYOL算法为例予以实施;基于Horovod框架进行大规模多机多卡训练及评测,以视频语义无监督学习PRP算法予以实施;所述训练包括环境配置、任务配置、通信配置、任务加速。本发明涉及的多机多卡大规模训练实验, batchsize之高,训练时间压缩之短,验证鹏城云脑I大科学装置的并行能力,拓展并行训练的集群规模,对于利用超大规模集群开展分布式训练具有指导意义。

云脑混合卡异构多队列申请机制



1. 一种大规模多机多卡预训练方法, 其特征在于, 所述大规模多机多卡预训练方法包括以下步骤:

步骤一, 基于slurm框架的多机多卡并行的无监督特征学习: 采用BYOL算法进行多机多卡部署, 进一步包括:

(1) 在云脑I上基于slurm进行多机部署, 由于控制节点, 即主节点不参与计算, 仅为控制节点申请CPU即可;

若服务器机型为DGX2, 则节点参数的配置情况, 包括:

以镜像的方式为每个节点配置运行环境;

控制节点的配置为: 控制节点任务不申请GPU, CPU核数申请6核, 内存申请100G, 并设置为主干任务; 子节点的配置为: 每个子节点任务申请16个GPU, CPU核数申请80核, 内存申请1T; 根据控制节点和子节点共享一台服务器, 总共使用的机器数量等于子节点的数量;

控制节点和子节点均采用IB/RDMA进行多机通信, 将内存的一半配置为共享内存;

配置启动命令和训练脚本, 开始运行并行训练任务;

(2) 基于debug模式配置多机slurm环境;

通过debug模式对DGX2的机器进行调试: 通过SSH进入到DGX2上执行的任务中; 若环境中没有安装slurm相关软件, 则进行安装, 若已经预先安装好slurm相关软件, 则通过slurmd -V和slurmd -C指令进行验证, 并查看CPU核数; 在云脑版slurm多机部署中, 修改masterip.txt和slaveip.txt, 在masterip.txt中添加主节点, 即控制节点的ip; 在slaveip.txt中添加子节点, 即计算节点的ip, 并修改slurm_autoconfig.sh脚本中的ControlMachine变量, 再执行bash slurm_autoconfig.sh脚本, 即可完成多机slurm环境的配置;

(3) 启动云脑I高速多机通信IB/RDMA

选用Nvidia NGC19.10作为基础镜像;

在训练脚本中指定IB网卡:

①os.environ['NCCL_IB_HCA'] = "mlx5_0";

②os.environ['NCCL_DEBUG'] = "INFO";

(4) 大规模任务的加速方案

采用如下措施对训练过程进行加速:

数据集存储加速: 采用开辟一个专用数据集存储空间并挂载内存的方式, 在不重启的情况下一直使用数据集, 通过加速数据读取达到加速训练的目的;

采用IB/RDMA进行多机通信, 通过加速训练中的多台机器之间的数据交互过程, 达到提高训练速度的目的;

采用Apex混合精度的方式进行训练, 占用显存相较单精度浮点型减少一半, 因此将batchsize扩大一倍;

采用适合大规模大batchsize情况的优化器, 通过优化器加速达到提高训练速度的目的;

针对slurm优化了CPU核数的分配, 在总核数一定的情况下, 尽可能提高CPU分配job效率;

(5) 基于步骤(1)~步骤(4), 将BYOL算法的slurm框架下多机多卡并行版本进行大规模

多机多卡的模型预训练;

(6) 利用单机多卡对BYOL算法的预训练模型进行评测;

步骤二, 基于Horovod框架的多机多卡并行的视频语义无监督学习: 采用PRP算法进行多机多卡部署, 所述视频语义无监督学习, 包括:

(1) 将Horovod框架部署到鹏城云脑I上;

通过镜像在节点上完成Horovod所需的软件环境的安装部署, 并设置好ssh免密登陆; 在启动任务时选择云脑I中已经安装好Horovod所需的软件的镜像; 在任务启动命令中加入ssh登录脚本的执行, 完成多机ssh免密登陆;

(2) 从部署环境后到开始训练的中间步骤, 除主节点可以申请GPU同时作为计算节点以外, 节点参数配置、启动云脑I高速多机通信IB/RDMA和大规模任务的加速方案同上;

(3) 将PRP算法改为Horovod框架下的多机多卡并行版本, 基于步骤(1)~步骤(2), 开展大规模多机多卡的模型预训练;

步骤三, 进行多机多卡混合机型训练。

2. 如权利要求1所述大规模多机多卡预训练方法, 其特征在于, 所述步骤一进一步包括:

使用N台DGX2共 $16 \times N$ 块V100, 采用无监督特征学习BYOL算法对Imagenet2012数据集进行训练, 获得预训练模型, 将训练时间从7天压缩到5小时6分钟;

设置一个CPU服务器作为主节点, 其他GPU服务器作为计算节点, 主节点和其中一台GPU服务器共享; 主节点和各子节点提交相应部署脚本。

3. 如权利要求1所述大规模多机多卡预训练方法, 其特征在于, 步骤(1)中, 所述多机免密配置脚本, 包括:

先生成秘钥, 包括私钥和公钥, 将秘钥拷贝至云脑平台的共享存储目录下, 并在该目录下新建ssh登录的shell脚本, 脚本内容为将秘钥分发至各个节点根目录.ssh路径下, 并修改相应权限; 该步骤方便后期通过云脑debug模式, 免密登录各任务所在机器进行调试。

4. 如权利要求2所述大规模多机多卡预训练方法, 其特征在于, 步骤三中, 所述多机多卡预训练, 包括:

在云脑I平台, 采用多队列申请方式实现不同机型的卡混合训练, 修改源文件进行重新挂载, 通过脚本编写去获得源文件的pod_id, 进行每个队列的ip和主机名的完善。

5. 一种应用如权利要求1~4任意一项所述大规模多机多卡预训练方法的大规模多机多卡预训练系统, 其特征在于, 所述大规模多机多卡预训练系统包括:

无监督特征学习模块, 用于采用BYOL算法进行多机多卡部署;

视频语义无监督学习模块, 用于采用PRP算法进行多机多卡部署;

多机多卡预训练模块, 用于进行多机多卡混合机型训练。

6. 一种计算机可读存储介质, 存储有如权利要求1~4任意一项所述大规模多机多卡预训练方法的计算机程序, 所述计算机程序被处理器执行时, 使得所述处理器执行如下步骤:

(1) 无监督特征学习: 采用BYOL算法进行多机多卡部署;

(2) 视频语义无监督学习: 采用PRP算法进行多机多卡部署;

(3) 多机多卡预训练: 进行多机多卡混合机型训练。

大规模多机多卡预训练方法、系统、设备及服务器集群

技术领域

[0001] 本发明属于分布式训练技术领域,尤其涉及一种大规模多机多卡预训练方法、系统、设备及服务器集群。

背景技术

[0002] 目前,针对我国对于AI开源开放共享创新平台的建设需求,鹏城实验室推出了鹏城云脑一期平台,鹏城云脑I是以英伟达GPU服务器为基础设施建设的一套大型集群系统,作为AI大科学装置用以支撑构造更好的AI生态,鹏城云脑I具备集群管理工具和资源调度平台,支持在GPU集群中运行AI任务。在智慧城市的建设升级过程中,数据量急剧增长,而且随着人工智能任务越来越复杂、越来越多样,模型规模也越来越大,目前实际应用中面临很多利用大规模数据对大模型进行训练的需求,利用多机多卡开展分布式训练是应对此类需求的必要途径。因此,基于鹏城云脑I进行大规模多机多卡分布式训练,能够显著地提高模型训练效率。

[0003] 就目前分布式训练使用的资源规模而言,OpenMMLab复现的BYOL算法公开的数据显示,最高仅用到128块GPU卡进行测试,batchsize最大为4096,目前国内外很少有单位能完成强大算力的大规模多机多卡运算,且超大数据集大batchsize训练时存在模型精度下降问题。另外,如何有效利用混合异构机器进行并行训练也是并行计算领域的一个难点,对于实际应用具有重要意义。因此,亟需一种新的大规模多机多卡预训练方法。

[0004] 通过上述分析,现有技术存在的问题及缺陷为:目前多机多卡数据训练中,现有技术很少有单位能完成强大算力的大规模多机多卡运算,且超大数据集大batchsize训练时存在模型精度下降问题;同时如何有效利用混合异构机器进行并行训练也是并行数据计算领域的一个难点。

[0005] 解决以上问题及缺陷的难度为:大规模多机多卡训练时的通信瓶颈和异常监测问题,大batchsize训练时如何选用合适的参数调整策略使得模型收敛且精度有所提升,保证稳定运行的同时确保算法性能。另外,随着技术发展和需求不同,不能保证资源池里都为同种类型机器,不同类型机器配置不同,怎样有效利用混合异构机器进行并行训练。

[0006] 解决以上问题及缺陷的意义为:更大规模集群的使用和模型成功训练极大地压缩了模型训练时间,提升了模型精度,为利用超大规模数据训练和使用通用大模型提供了技术途径,在保证算法性能的同时可以支撑更多的下游任务;混合异构机器的并行训练可提升资源利用率,进一步提升并行训练规模。

发明内容

[0007] 针对现有技术存在的问题,本发明提供了一种大规模多机多卡预训练方法、系统、设备及服务器集群,尤其涉及一种基于鹏城云脑I的大规模多机多卡(GPU)预训练方法、系统、设备及服务器集群。

[0008] 本发明是这样实现的,一种大规模多机多卡预训练方法,包括:

- [0009] 在多个服务器上部署多机多卡,进行同构机型和异构混合机型的多机多卡并行;
- [0010] 基于slurm框架进行大规模多机多卡训练及评测,以无监督特征学习BYOL算法为例予以实施;
- [0011] 基于Horovod框架进行大规模多机多卡训练及评测,以视频语义无监督学习PRP算法予以实施;
- [0012] 所述训练包括环境配置、任务配置、通信配置、任务加速等。
- [0013] 具体包括:
- [0014] 步骤一,无监督特征学习:采用BYOL算法进行多机多卡部署;
- [0015] 步骤二,视频语义无监督学习:采用PRP算法进行多机多卡部署;
- [0016] 步骤三,多机多卡预训练:进行多机多卡混合机型训练。
- [0017] 进一步,步骤一中,进一步包括:
- [0018] 使用N台DGX2共 $16 \times N$ 块V100,采用无监督特征学习BYOL算法对Imagenet2012数据集进行训练,获得预训练模型,将训练时间从7天压缩到5小时6分钟,所述方案包括:设置一个CPU服务器作为主节点,其他GPU服务器作为计算节点,主节点和其中一台GPU服务器共享;主节点和各子节点提交相应部署脚本。
- [0019] 进一步,步骤一中,进一步包括:
- [0020] (1) 在云脑I上基于slurm进行多机部署,由于控制节点,即主节点不参与计算,仅为控制节点申请CPU即可;
- [0021] 若服务器机型为DGX2,则所述节点参数的配置情况,包括:
- [0022] 1) 以镜像的方式为每个节点配置运行环境;
- [0023] 2) 控制节点的配置为:控制节点任务不申请GPU,CPU核数申请6核,内存申请100G,并设置为主干任务;子节点的配置为:每个子节点任务申请16个GPU,CPU核数申请80核,内存申请1T;根据控制节点和子节点共享一台服务器,总共使用的机器数量等于子节点的数量;
- [0024] 3) 控制节点和子节点均采用IB/RDMA进行多机通信,将内存的一半配置为共享内存;
- [0025] 4) 配置启动命令和训练脚本,开始运行并行训练任务。
- [0026] (2) 基于debug模式配置多机slurm环境。
- [0027] 通过debug模式对DGX2的机器进行调试:通过SSH进入到DGX2上执行的任务中;若环境中没有安装slurm相关软件,则进行安装,若已经预先安装好slurm相关软件,则通过slurmd-V和slurmd-C指令进行验证,并查看CPU核数;在云脑版slurm多机部署中,修改masterip.txt和slaveip.txt,在masterip.txt中添加主节点,即控制节点的ip;在slaveip.txt中添加子节点,即计算节点的ip,并修改slurm_autoconfig.sh脚本中的ControlMachine变量,再执行bash slurm_autoconfig.sh脚本,即可完成多机slurm环境的配置。
- [0028] (3) 启动云脑I高速多机通信IB/RDMA
- [0029] 1) 选用Nvidia NGC19.10作为基础镜像,镜像链接为:https://docs.nvidia.com/deeplearning/frameworks/pytorch-release-notes/rel_19-10.html#rel_19-10;
- [0030] 2) 在训练脚本中指定IB网卡;

[0031] ①os.environ['NCCL_IB_HCA']="mlx5_0";

[0032] ②os.environ['NCCL_DEBUG']="INFO".

[0033] (4) 大规模任务的加速方案

[0034] 采用如下措施对训练过程进行加速:

[0035] 1) 数据集存储加速:采用开辟一个专用数据集存储空间并挂载内存的方式,在不重启的情况下一直使用数据集,通过加速数据读取达到加速训练的目的;

[0036] 2) 采用IB/RDMA进行多机通信,通过加速训练中的多台机器之间的数据交互过程,达到提高训练速度的目的;

[0037] 3) 采用Apex混合精度的方式进行训练,占用显存相较单精度浮点型减少一半,因此也可以将batchsize扩大一倍;

[0038] 4) 采用适合大规模大batchsize情况的优化器,例如lars、lamb、yogi等,通过优化器加速达到提高训练速度的目的;

[0039] 5) 针对slurm优化了CPU核数的分配,在总核数一定的情况下,尽可能提高CPU分配job效率。

[0040] (5) 基于步骤(1)~步骤(4),将BYOL算法的slurm框架下多机多卡并行版本进行大规模多机多卡的模型预训练。

[0041] (6) 利用单机多卡对BYOL算法的预训练模型进行评测。

[0042] BYOL算法作者以resnet50为基网、总batchsize为4096训练200轮得到预训练模型,并在预训练模型的基础上使用单机8卡、总batchsize为256进行ImageNet Linear Classification任务的评测,在评测任务训练的100轮中,验证集上最高的top1-accuracy为67.10。

[0043] 所述评测使用的预训练模型为:使用8台DGX2共128块V100、总batchsize为12288、基网resnet101训练200轮得到的预训练模型;在预训练模型的基础上采用单机16卡、总batchsize为2048进行ImageNet Linear Classification任务的评测,在评测任务训练的100轮中,验证集上最高的top1-accuracy为69.294。

[0044] 进一步,步骤二中,所述视频语义无监督学习,包括:

[0045] (1) 将Horovod框架部署到鹏城云脑I上。

[0046] 通过镜像在节点上完成Horovod所需的软件环境的安装部署,并设置好ssh免密登陆;在启动任务时选择云脑I中已经安装好Horovod所需的软件的镜像;在任务启动命令中加入ssh登录脚本的执行,完成多机ssh免密登陆。

[0047] (2) 从部署环境后到开始训练的中间步骤,除主节点可以申请GPU同时作为计算节点以外,节点参数配置、启动云脑I高速多机通信IB/RDMA和大规模任务的加速方案同上。

[0048] (3) 将PRP算法改为Horovod框架下的多机多卡并行版本,基于步骤(1)~步骤(2),开展大规模多机多卡的模型预训练。

[0049] 进一步,步骤(1)中,所述多机免密配置脚本,包括:

[0050] 先生成密钥,包括私钥和公钥,将密钥拷贝至云脑平台的共享存储目录下,并在该目录下新建ssh登录的shell脚本,脚本内容为将密钥分发至各个节点根目录.ssh路径下,并修改相应权限;该步骤方便后期通过云脑debug模式,免密登录各任务所在机器进行调试。

[0051] 进一步,步骤三中,所述多机多卡预训练,包括:

[0052] 在云脑I平台,采用多队列申请方式实现不同机型的卡混合训练,修改源文件进行重新挂载,通过脚本编写去获得源文件的pod_id,进行每个队列的ip和主机名的完善,这样重新挂载进去之后,每个队列中的/etc/hosts就会有全部队列的ip和主机名,保证机器间可通信。

[0053] 本发明的另一目的在于提供一种应用所述大规模多机多卡预训练方法的大规模多机多卡预训练系统,所述大规模多机多卡预训练系统包括:

[0054] 无监督特征学习模块,用于采用BYOL算法进行多机多卡部署;

[0055] 视频语义无监督学习模块,用于采用PRP算法进行多机多卡部署;

[0056] 多机多卡预训练模块,用于进行多机多卡混合机型训练。

[0057] 本发明的另一目的在于提供一种计算机设备,所述计算机设备包括存储器和处理器,所述存储器存储有计算机程序,所述计算机程序被所述处理器执行时,使得所述处理器执行如下步骤:

[0058] (1) 无监督特征学习:采用BYOL算法进行多机多卡部署;

[0059] (2) 视频语义无监督学习:采用PRP算法进行多机多卡部署;

[0060] (3) 多机多卡预训练:进行多机多卡混合机型训练。

[0061] 本发明的另一目的在于提供一种计算机可读存储介质,存储有计算机程序,所述计算机程序被处理器执行时,使得所述处理器执行如下步骤:

[0062] (1) 无监督特征学习:采用BYOL算法进行多机多卡部署;

[0063] (2) 视频语义无监督学习:采用PRP算法进行多机多卡部署;

[0064] (3) 多机多卡预训练:进行多机多卡混合机型训练。

[0065] 本发明的另一目的在于提供一种信息数据处理服务器集群,所述信息数据处理服务器集群用于实现所述大规模多机多卡预训练系统。

[0066] 结合上述的所有技术方案,本发明所具备的优点及积极效果为:本发明提供的大规模多机多卡预训练方法,采用25台DGX2共400块V100完成了测试,batchsize最高为76800,目前国内外很少有单位能完成如此强大算力的大规模多机多卡运算。本发明中涉及的多机多卡大规模训练实验,batchsize之高,训练时间压缩之短,既验证了鹏城云脑I大科学装置的并行能力,又进一步拓展了并行训练的集群规模,对于利用超大规模集群开展分布式训练的可行性和具体实施办法具有极大的指导意义,同时,模型精度在预训练之后进行的下游任务评测上也表现出很好的性能。涉及的异构混合机型训练方案,为基于混合异构平台开展大规模的模型训练提供了一种实现思路。

[0067] 本发明提供的无监督特征学习应用实例的实施情况如下:

[0068] (1) 对标工作为以resnet50为基网、总batchsize为4096训练200轮得到预训练模型,并在预训练模型的基础上进行ImageNetLinearClassification任务的评测,在评测任务中,总共训练100轮,采用单机8卡以及总batchsize为256,在验证集上得到的最高top1精度为67.10;

[0069] (2) 本发明以resnet101为基网、总batchsize为12288训练200轮得到预训练模型,并在预训练模型的基础上进行ImageNetLinearClassification任务的评测,在评测任务中,总共训练100轮,采用单机16卡、总batchsize为2048进行测试,在验证集上得到的最高

top1精度为69.294;

[0070] (3) 采用和作者一致的参数,本发明在验证集上得到的最高top1精度为67.068,而作者公开的最高数据为67.10,两者基本一致。

[0071] (4) 对于视频语义无监督学习的应用,由于PRP算法暂无多机多卡程序,因此,对原单机多卡程序进行了修改,完成多机多卡训练和测试。

[0072] 本发明基于鹏城云脑I平台,提出并实现了大规模多机多卡分布式训练方法,成功实现了同构机型和异构混合机型的多机多卡并行。分别基于slurm框架和Horovod框架实现了BYOL算法和PRP算法的大规模多机多卡训练及评测的全过程,包括环境配置、任务配置、通信配置、任务加速等,极大地缩短了大模型在大数据集上的训练时间,同时也进一步扩大了参与训练的机器规模,两套多机多卡大规模运算算法的训练时间压缩均达到预期目标。

[0073] 同机型多机多卡并行的完成情况为:

[0074] (1) 使用25台DGX2共400块V100,部署slurm框架,配置多机并行环境,结合多种加速策略,采用无监督特征学习BYOL算法,对Imagenet2012数据集进行训练,获得预训练模型,将训练时间从7天压缩到5小时6分钟,完成预期目标;

[0075] (2) 使用14台DGX2共224块V100,部署Horovod框架,配置多机并行环境,结合多种加速策略,采用视频语义无监督学习PRP算法,对Kinetics400数据集进行训练,将训练时间从10天压缩到11小时42分钟,完成预期目标。

[0076] (3) 对于混合机型,如云脑I上的DGX1、DGX2、AGX等机型,实现了以上两个算法的多机型的混合训练。

[0077] 本发明的关键点是:

[0078] (1) 基于鹏城云脑I平台,提出了大规模多机多卡训练方法,成功实现了同构机型和异构混合机型两种情况下的多机多卡并行,分别基于slurm框架和Horovod框架实现了BYOL算法和PRP算法的大规模多机多卡训练及评测的全过程,极大地缩短了大模型在大数据集上的训练时间;

[0079] (2) 针对大batchsize情况下训练loss不收敛的问题,采用适当的优化器如lars、lamb、yogi等,以及合适的参数调整策略,解决了超大数据集大batchsize训练时的模型精度下降问题,并进一步拓展了参与训练的机器数量和batchsize大小,提升了模型精度;

[0080] (3) 采用IB/RDMA进行多机通信,并基于Apex的混合精度加速进行训练,进一步加快了训练速度和减少了资源消耗;

[0081] (4) 提出了一种基于DGX2、DGX1和AGX多种机型混合异构平台的分布式训练方法,便于在资源受限的情况下充分利用已有异构设备开展多机多卡并行训练,也为后续基于混合异构平台开展更大规模的模型训练提供了思路和参考。

附图说明

[0082] 为了更清楚地说明本发明实施例的技术方案,下面将对本发明实施例中所需要使用的附图做简单的介绍,显而易见地,下面所描述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。

[0083] 图1是本发明实施例提供的大规模多机多卡预训练方法流程图。

- [0084] 图2是本发明实施例提供的云脑I混合机型训练多队列申请机制示意图。
- [0085] 图3是本发明实施例提供的大规模多机多卡预训练系统结构框图；
- [0086] 图中：1、无监督特征学习模块；2、视频语义无监督学习模块；3、多机多卡预训练模块。

具体实施方式

[0087] 为了使本发明的目的、技术方案及优点更加清楚明白，以下结合实施例，对本发明进行进一步详细说明。应当理解，此处所描述的具体实施例仅仅用以解释本发明，并不用于限定本发明的保护范围。

[0088] 针对现有技术存在的问题，本发明提供了一种大规模多机多卡预训练方法、系统、设备及服务器集群，下面结合附图对本发明作详细的描述。

[0089] 如图1所示，本发明实施例提供的大规模多机多卡预训练方法包括以下步骤：

[0090] S101，无监督特征学习：采用BYOL算法进行多机多卡部署；

[0091] S102，视频语义无监督学习：采用PRP算法进行多机多卡部署；

[0092] S103，多机多卡预训练：进行云脑I混合机型训练。

[0093] 本发明实施例提供的云脑I混合机型训练多队列申请机制示意图如图2所示。

[0094] 如图3所示，本发明实施例提供的大规模多机多卡预训练系统包括：

[0095] 无监督特征学习模块1，用于采用BYOL算法进行多机多卡部署；

[0096] 视频语义无监督学习模块2，用于采用PRP算法进行多机多卡部署；

[0097] 多机多卡预训练模块3，用于进行云脑I混合机型训练。

[0098] 下面结合具体实施例对本发明的技术方案作进一步描述。

[0099] 实施例

[0100] 1、使用鹏城云脑I进行多机多卡（GPU）并行训练；

[0101] 对于AI开源开放共享创新平台的建设需求，鹏城实验室推出了鹏城云脑一期平台，鹏城云脑I是以英伟达GPU服务器为基础设施建设的一套大型集群系统，作为AI大科学装置用以支撑构造更好的AI生态，鹏城云脑I具备集群管理工具和资源调度平台，支持在GPU集群中运行AI任务。在智慧城市的建设升级过程中，数据量急剧增长，而且随着人工智能任务越来越复杂、越来越多样，模型规模也越来越大，目前实际应用中面临很多利用大规模数据对大模型进行训练的需求，利用多机多卡开展分布式训练是应对此类需求的必要途径。因此，基于鹏城云脑I进行大规模多机多卡分布式训练，能够充分挖掘其集群优势和并行化能力，显著地提高模型训练效率，提升模型训练效果，充分发挥大科学装置的作用。

[0102] 基于鹏城云脑I的大规模多机多卡训练任务主要目标有两个：第一个为无监督特征学习，采用BYOL算法对Imagenet2012数据集（约120多万图片）进行训练，将训练时间从7天压缩到0.25天；第二个为视频语义无监督学习，采用PRP算法对Kinetics 400数据集（约5000万帧短视频）进行训练，将训练时间从10天压缩到0.5天。

[0103] 本发明提供的大规模多机多卡预训练方法，采用25台DGX2共400块V100完成了测试，batchsize最高为76800，目前国内外很少有单位能完成如此强大算力的大规模多机多卡运算。本发明中涉及的多机多卡大规模训练实验，batchsize之高，训练时间压缩之短，既验证了鹏城云脑I大科学装置的并行能力，又进一步拓展了并行训练的集群规模，对于利用

超大规模集群开展分布式训练的可行性和具体实施办法具有极大的指导意义,同时,模型精度在预训练之后进行的下游任务评测上也表现出很好的性能。涉及的异构混合机型训练方案,为基于混合异构平台开展大规模的模型训练提供了一种实现思路。

[0104] 本发明提供的无监督特征学习应用实例的实施情况如下:

[0105] (1) 对标工作为以resnet50为基网、总batchsize为4096训练200轮得到预训练模型,并在预训练模型的基础上进行ImageNetLinearClassification任务的评测,在评测任务中,总共训练100轮,作者采用单机8卡以及总batchsize为256,在验证集上得到的最高top1精度为67.10;

[0106] (2) 本发明以resnet101为基网、总batchsize为12288训练200轮得到预训练模型,并在预训练模型的基础上进行ImageNetLinearClassification任务的评测,在评测任务中,总共训练100轮,采用单机16卡、总batchsize为2048进行测试,在验证集上得到的最高top1精度为69.294;

[0107] (3) 采用和作者一致的参数,本发明在验证集上得到的最高top1精度为67.068,而作者公开的最高数据为67.10,两者基本一致。

[0108] (4) 对于视频语义无监督学习的应用,由于PRP算法暂无多机多卡程序,因此,对原单机多卡程序进行了修改,完成多机多卡训练和测试。

[0109] 本发明基于鹏城云脑I平台,提出并实现了大规模多机多卡分布式训练方法,成功实现了同构机型和异构混合机型的多机多卡并行。分别基于slurm框架和Horovod框架实现了BYOL算法和PRP算法的大规模多机多卡训练及评测的全过程,包括环境配置、任务配置、通信配置、任务加速等,极大地缩短了大模型在大数据集上的训练时间,同时也进一步扩大了参与训练的机器规模,两套多机多卡大规模运算算法的训练时间压缩均达到预期目标。

[0110] 同机型多机多卡并行的完成情况为:

[0111] (1) 使用25台DGX2共400块V100,部署slurm框架,配置多机并行环境,结合多种加速策略,采用无监督特征学习BYOL算法,对Imagenet2012数据集进行训练,获得预训练模型,将训练时间从7天压缩到5小时6分钟,完成预期目标。

[0112] (2) 使用14台DGX2共224块V100,部署Horovod框架,配置多机并行环境,结合多种加速策略,采用视频语义无监督学习PRP算法,对Kinetics400数据集进行训练,将训练时间从10天压缩到11小时42分钟,完成预期目标。

[0113] (3) 对于混合机型,如云脑I上的DGX1、DGX2、AGX等机型,实现了以上两个算法的多机型的混合训练。

[0114] 2、本发明是在多个服务器上采用多机多卡进行运行。

[0115] 目前鹏城云脑I上有三种服务器机型,DGX1、DGX2、和AGX,具体参数配置如表1所示。

[0116] 表1服务器配置

[0117]

机型	IB网卡	功率	CPU核数	内存	V100
DGX2	每台8个	350W	96核	1.5T	单机16卡
DGX1	每台2个	163W	80核	0.5T	单机8卡
AGX	每台2个	300W	96核	1.5T	单机8卡

[0118] (2.1) 无监督特征学习:BYOL算法多机多卡部署方案

[0119] 使用N台DGX2共 $16 \times N$ 块V100,采用无监督特征学习BYOL算法对Imagenet2012数据集进行训练,获得预训练模型,将训练时间从7天压缩到5小时06分钟。大致方案为:设置一个CPU服务器作为主节点,其他GPU服务器作为计算节点,主节点可以和其中一台GPU服务器共享;主节点和各子节点提交相应部署脚本。

[0120] 具体部署流程如下:

[0121] 101.在云脑I上基于slurm进行多机部署,由于控制节点(主节点)不参与计算,仅需要为控制节点申请CPU即可;

[0122] 以DGX2机器为例,来说明节点参数的详细配置情况。

[0123] (1)以镜像的方式为每个节点配置运行环境。

[0124] (2)控制节点的配置为:控制节点任务不申请GPU,CPU核数申请6核,内存申请100G,并设置为主干任务;子节点的配置为:每个子节点任务申请16个GPU,CPU核数申请80核,内存申请1T。根据控制节点和子节点可以共享一台服务器,总共使用的机器数量等于子节点的数量。

[0125] (3)控制节点和子节点均采用IB/RDMA进行多机通信,将内存的一半配置为共享内存。

[0126] (4)配置启动命令和训练脚本,开始运行并行训练任务。

[0127] 102.基于debug模式配置多机slurm环境。

[0128] 通过debug模式对DGX2的机器进行调试。首先,通过SSH进入到DGX2上执行的任务中;然后,若环境中没有安装slurm相关软件,则进行安装,若已经预先安装好slurm相关软件,则可通过slurmd-V和slurmd-C指令进行验证,并查看CPU核数;最后,在云脑版slurm多机部署中,需修改masterip.txt和slaveip.txt,在masterip.txt中添加主节点(控制节点)的ip,在slaveip.txt中添加子节点(计算节点)的ip,并修改slurm_autoconfig.sh脚本中的ControlMachine变量,再执行bash slurm_autoconfig.sh脚本,即可完成多机slurm环境的配置。

[0129] 103.启动云脑I高速多机通信IB/RDMA

[0130] (1)选用Nvidia NGC19.10作为基础镜像,镜像链接为:https://docs.nvidia.com/deeplearning/frameworks/pytorch-release-notes/rel_19-10.html#rel_19-10。

[0131] (2)在训练脚本中指定IB网卡:

[0132]

```
os.environ['NCCL_IB_HCA'] = "mlx5_0"
os.environ['NCCL_DEBUG'] = "INFO"
```

[0133] 104.大规模任务的加速方案

[0134] 采用如下措施对训练过程进行加速。

[0135] (1)数据集存储加速:采用开辟一个专用数据集存储空间并挂载内存的方式,可以在不重启的情况下一直使用数据集,通过加速数据读取达到加速训练的目的;

[0136] (2)采用IB/RDMA进行多机通信,通过加速训练中的多台机器之间的数据交互过程,达到提高训练速度的目的;

[0137] (3)采用Apex混合精度的方式进行训练,占用显存相较单精度浮点型减少一半,因此也可以将batchsize扩大一倍;

[0138] (4) 采用适合大规模大batchsize情况的优化器,例如lars、lamb、yogi等,通过优化器加速达到提高训练速度的目的;

[0139] (5) 针对slurm优化了CPU核数的分配,在总核数一定的情况下,尽可能提高CPU分配job效率。

[0140] 105. 基于上述步骤,将BYOL算法的slurm框架下多机多卡并行版本进行大规模多机多卡的模型预训练,具体运行结果如下:

[0141] 表2 BYOL算法多机多卡训练结果

[0142]

集群配置	数据存储	IB 网卡	Batchsize	Apex fp16 加速	cpus-per-task	训练 时间
单机 16 卡	硬盘	\	96*16= 1536	否	\	60h 36m
			192*16= 3072	是		55h 12m
	内存	\	96*16= 1536	否	\	59h 57m
			96*16= 1536	是		39h 58m
			192*16= 3072			39h 15m
2 机 32 卡	内存	1	96*32= 3072	否	4	33h 12m
7 机 112 卡	内存	4	96*112= 11232	否	4	11h 51m
8 机 128 卡	内存	1	96*128= 12288	否	1	28h 30m
					2	17h 06m
					3	13h 30m
					4	11h 06m
		8	96*128=	否	4	10h 04m

[0143]

			12288	是		8h 48m
			192*128=24576			
12 机 192 卡	内存	1	96*192=18432	否	4	8h 42m
14 机 224 卡	内存	1	96*224=21504	否	4	7h 32m
16 机 256 卡	硬盘	1	96*256=24576	否	4	8h 24m
	内存					7h 54m
25 机 400 卡	硬盘	1	96*400=38400	否	4	6h 46m
	内存	1	96*400=38400	否		5h 43m
		1	96*400=38400	是		5h 18m
		1	192*400=76800	是		5h 06m
29 机 464 卡	硬盘	接满	96* 464=44544	否	4	6h

[0144] 106. 利用单机多卡对BYOL算法的预训练模型进行评测。

[0145] BYOL算法作者以resnet50为基网、总batchsize为4096训练200轮得到预训练模型,并在预训练模型的基础上使用单机8卡、总batchsize为256进行ImageNet Linear Classification任务的评测,在评测任务训练的100轮中,验证集上最高的top1-accuracy为67.10。

[0146] 本发明评测使用的预训练模型为:使用8台DGX2共128块V100、总batchsize为12288、基网resnet101(比resnet50参数大)训练200轮得到的预训练模型;并在预训练模型的基础上采用单机16卡、总batchsize为2048进行ImageNet Linear Classification任务的评测,在评测任务训练的100轮中,验证集上最高的top1-accuracy为69.294,与原作者相比,模型性能有一定的提升。

[0147] 表3 ImageNet Linear Classification任务评测结果

[0148]

方法	基网	训练次数	预训练模型训练的batchsize	评测任务的batchsize	验证集Top1-accuracy
BYOL 作者	Resnet50	相同	4096	256	67.10
本专利	Resnet101		12288	256	68.70
				512	68.994

[0149]

				1024	68.984
				2048	69.294
在预训练模型评测任务中，第一次评测除了基网以及预训练时的 batchsize 不一样之外，其余在评测任务中的指标均和作者保持一致，验证集的 Top1-accuracy 由 67.10 上升到 68.70。					

[0150] (2.2)PRP算法多机多卡部署方案

[0151] 201.将Horovod框架部署到鹏城云脑I上。

[0152] 首先,通过镜像在节点上完成Horovod所需的软件环境的安装部署,并设置好ssh免密登陆。此处启动任务时选择云脑I中已经安装好Horovod所需的软件的镜像即可,其中nccl版本选择了2.7.8。

[0153] 然后,在任务启动命令中加入ssh登录脚本的执行,完成多机ssh免密登陆。其中多机免密配置脚本,执行步骤如下:先生成密钥(包括私钥和公钥),将密钥拷贝至云脑平台的共享存储目录下,并在该目录下新建ssh登录的shell脚本,脚本内容为将密钥分发至各个节点根目录.ssh路径下,并修改相应权限。该步骤方便后期通过云脑debug模式,免密登录各任务所在机器进行调试。

[0154] 202.从部署环境后到开始训练的中间步骤,除主节点可以申请GPU同时作为计算节点以外,节点参数配置、启动云脑I高速多机通信IB/RDMA和大规模任务的加速方案同上。

[0155] 203.将PRP算法改为Horovod框架下的多机多卡并行版本,基于上述步骤,开展大规模多机多卡的模型预训练,具体运行结果如下:

[0156] 表4 PRP算法多机多卡训练结果

[0157]

集群配置	Batchsize	IB网卡	数据存储	训练时间
8机128卡	36*128=4608	1根线	硬盘	20h 30m
12机192卡	32*192=6144	1根线	硬盘	14h 7m
13机208卡	32*208=6656	1根线	内存	12h 25m
14机224卡	36*224=8064	1根线	内存	11h 42m

[0158] (2.3)云脑I混合机型训练方案

[0159] 在云脑I平台,采用多队列申请方式实现不同机型的卡混合训练,该方式需要解决的就是每个队列下的/etc/hosts修改问题,由于etc/hosts实际上是外部写好的源文件挂载进去的,修改了源文件进行重新挂载就能解决etc/hosts无法挂载问题,通过脚本编写去获得源文件的pod_id,进行每个队列的ip和主机名的完善,这样重新挂载进去之后,每个队列中的/etc/hosts就会有全部队列的ip和主机名,保证机器间可通信。

[0160] 具体流程图如图2所示。

[0161] 举例说明多队列资源申请步骤:使用5台DGX2+1台DGX1+1台AGX实现并行。队列1,申请1台DGX2,启动1个任务,设置为主节点;队列2,申请4台DGX2,启动4个任务,设置为子节点;队列3,申请1台AGX,启动1个任务,设置为子节点;队列4,申请1台DGX1,启动1个任务,设置为子节点。任务的配置参数以及训练加速方案同上。

[0162] 本发明的关键点是:

[0163] (1)基于鹏城云脑I平台,提出了大规模多机多卡训练方法,成功实现了同构机型和异构混合机型两种情况下的多机多卡并行,分别基于slurm框架和Horovod框架实现了BYOL算法和PRP算法的大规模多机多卡训练及评测的全过程,极大地缩短了大模型在大数据集上的训练时间;

[0164] (2)针对大batchsize情况下训练loss不收敛的问题,采用适当的优化器如lars、lamb、yogi等,以及合适的参数调整策略,解决了超大数据集大batchsize训练时的模型精度下降问题,并进一步拓展了参与训练的机器数量和batchsize大小,提升了模型精度;

[0165] (3) 采用IB/RDMA进行多机通信,并基于Apex的混合精度加速进行训练,进一步加快了训练速度和减少了资源消耗;

[0166] (4) 提出了一种基于DGX2、DGX1和AGX多种机型混合异构平台的分布式训练方法,便于在资源受限的情况下充分利用已有异构设备开展多机多卡并行训练,也为后续基于混合异构平台开展更大规模的模型训练提供了思路 and 参考。

[0167] prp算法的单机多卡代码链接:<https://github.com/yuanyao366/PRP>;

[0168] prp算法的论文链接:<https://arxiv.org/abs/2006.11476>;

[0169] 本发明的其中一个贡献在于将prp算法修改为多机多卡版本,并在鹏城云脑1上运行。

[0170] 在本发明的描述中,除非另有说明,“多个”的含义是两个或两个以上;术语“上”、“下”、“左”、“右”、“内”、“外”、“前端”、“后端”、“头部”、“尾部”等指示的方位或位置关系为基于附图所示的方位或位置关系,仅是为了便于描述本发明和简化描述,而不是指示或暗示所指的装置或元件必须具有特定的方位、以特定的方位构造和操作,因此不能理解为对本发明的限制。此外,术语“第一”、“第二”、“第三”等仅用于描述目的,而不能理解为指示或暗示相对重要性。

[0171] 在上述实施例中,可以全部或部分地通过软件、硬件、固件或者其任意组合来实现。当使用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。在计算机上加载或执行所述计算机程序指令时,全部或部分地产生按照本发明实施例所述的流程或功能。所述计算机可以是通用计算机、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(例如同轴电缆、光纤、数字用户线(DSL)或无线(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)。所述计算机可读取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成的服务器、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(例如固态硬盘Solid State Disk(SSD))等。

[0172] 以上所述,仅为本发明的具体实施方式,但本发明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发明揭露的技术范围内,凡在本发明的精神和原则之内所作的任何修改、等同替换和改进等,都应涵盖在本发明的保护范围之内。

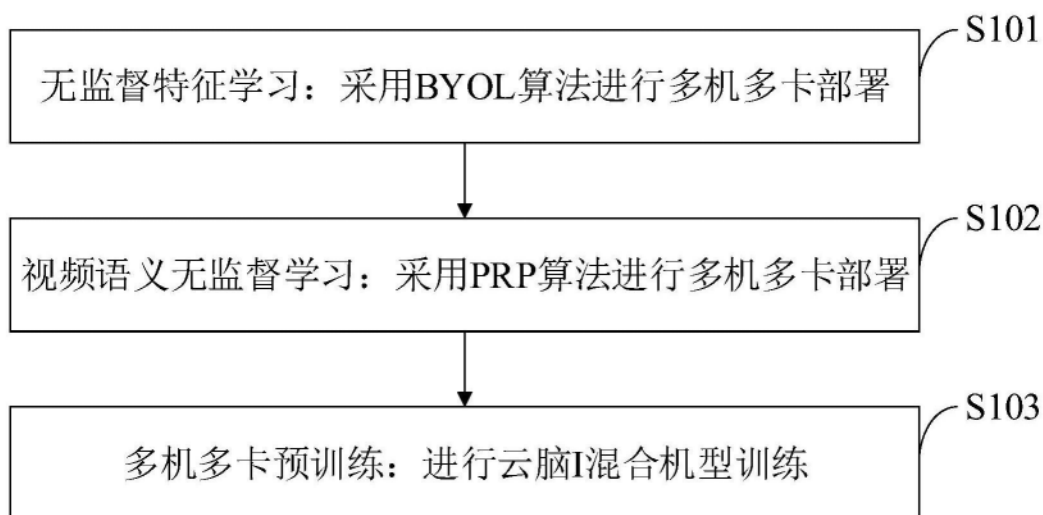


图1

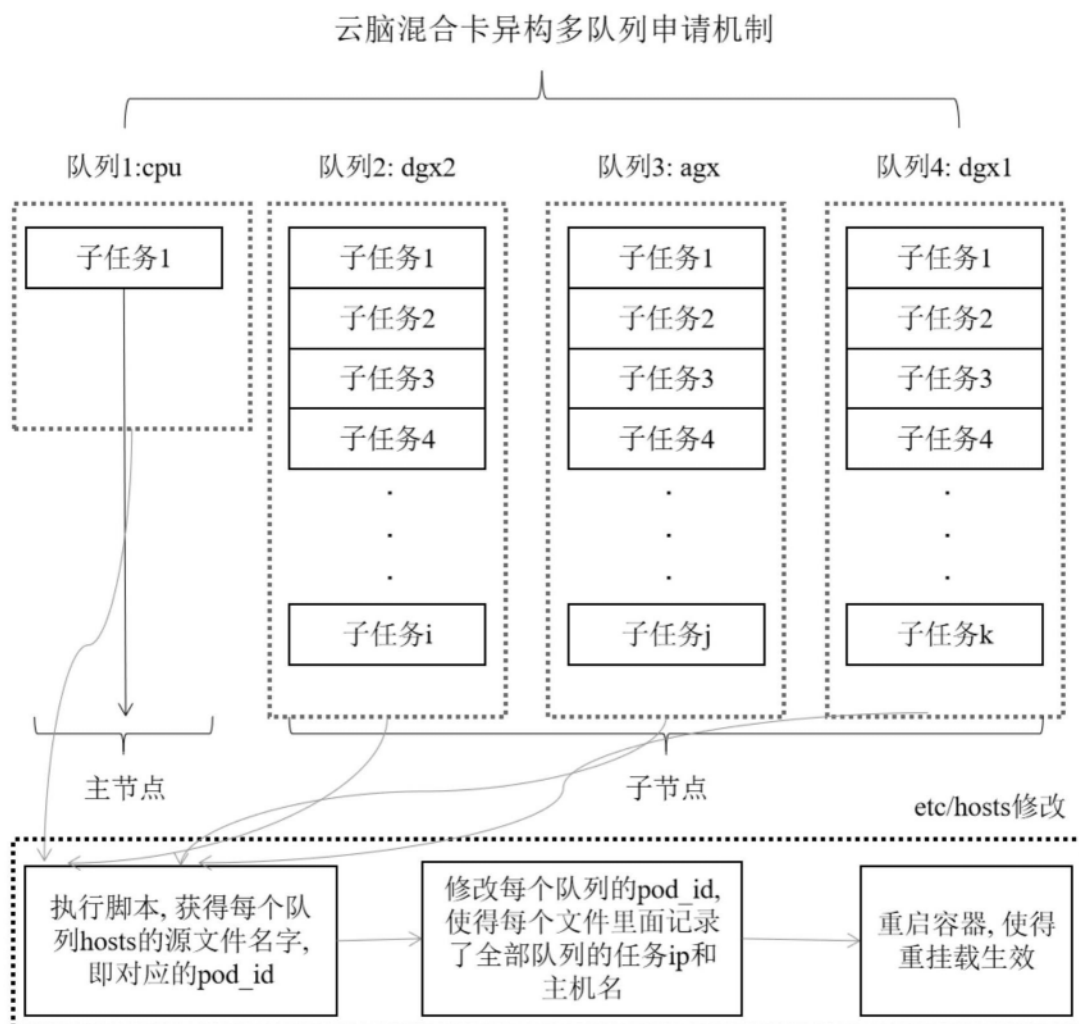


图2

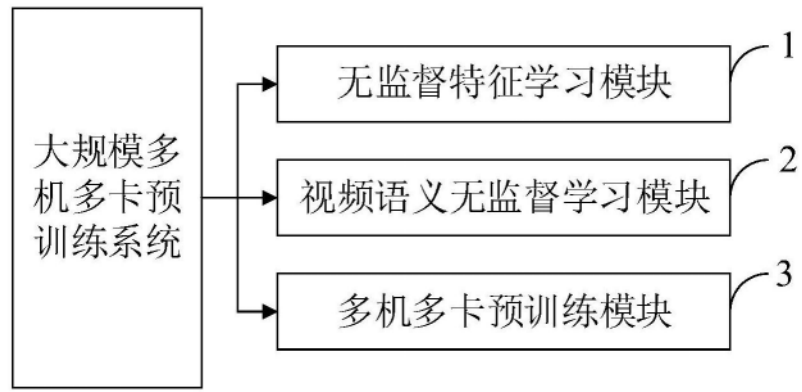


图3