(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113687936 B
(45)授权公告日 2024.07.30

(21)申请号 202111082553.2

(22)申请日 2021.09.15

(65)同一申请的已公布的文献号
申请公布号 CN 113687936 A

(43)申请公布日 2021.11.23

(66)本国优先权数据
202110604769.4 2021.05.31 CN

(73)专利权人 杭州云栖智慧视通科技有限公司
地址 310000 浙江省杭州市西湖区转塘科技经济区块16号2幢401室

(72)发明人 姜枞聪 樊一超 曹航 李冠华

(74)专利代理机构 杭州信与义专利代理有限公司 33450
专利代理师 马育妙

(51)Int.Cl.
G06F 9/48(2006.01)

(56)对比文件
CN 103955409 A,2014.07.30
CN 106469352 A,2017.03.01

审查员 刘梅

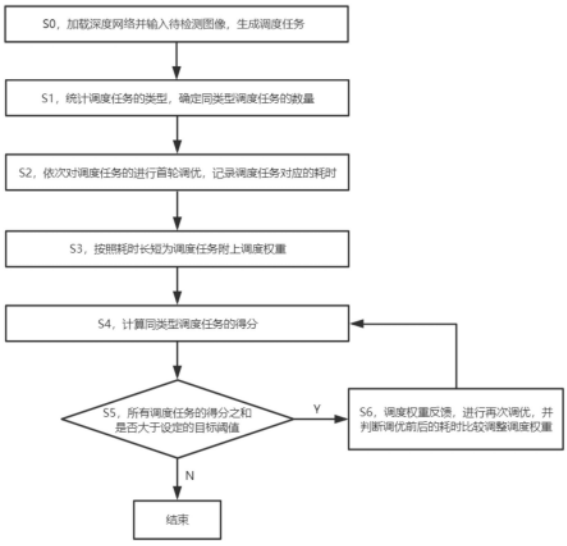
权利要求书2页 说明书4页 附图1页

(54)发明名称

在TVM中加速调优收敛的调度方法、存储介质和电子设备

(57)摘要

本发明公开了一种在TVM中加速调优收敛的调度方法,包括以下步骤:统计调度任务的类型,确定同类型调度任务的数量;首轮遍历调优,记录调度任务对应的耗时;按照耗时长短为调度任务附上调度权重;计算同类型调度任务的得分;以所有调度任务的得分之和作为目标函数值与设定的目标阈值作比较,若目标函数值大于目标阈值,则进行调度权重反馈;调度权重反馈,进行再次调优,并判断调优前后的耗时比较调整调度权重;本发明还同步公开一种、存储介质和电子设备。本发明对调优后的耗时与调优前进行比较,并进行相应的调度权重奖励与惩罚,使调度能够更加快速地收敛。



CN 113687936 B

1.一种在TVM中加速调优收敛的调度方法,其特征在于,包括以下步骤:

S1,统计调度任务的类型,确定每一类型调度任务的数量;

S2,首轮遍历调优,依次对调度任务进行首轮调优,记录调度任务对应的耗时;

S3,按照耗时长短为调度任务附上调度权重,耗时越长调度权重越重;

S4,计算耗时、每一类型调度任务的数量、调度权重三者的乘积为该类型调度任务的得分;

S5,以所有调度任务的得分之和作为目标函数值与设定的目标阈值作比较,若目标函数值大于目标阈值,则跳至S6进行调度权重反馈,否则结束调优;

S6,调度权重反馈,对得分最大的该类型调度任务进行再次调优,并判断调优前后的耗时,若调优后的耗时比调优前的耗时短,则加重其调度权重作为奖励,否则减轻其调度权重作为惩罚;跳回至S4。

2.根据权利要求1所述的在TVM中加速调优收敛的调度方法,其特征在于,所述S2中,首轮遍历调优的具体内容为:初始化首轮遍历索引;对所述首轮遍历索引累加1,并依次对所述首轮遍历索引对应的调度任务进行调优直至所有调度任务完成首轮遍历调优。

3.根据权利要求1所述的在TVM中加速调优收敛的调度方法,其特征在于,所述S6中,调度权重反馈的具体内容为:

S61,选取得分最大的该类型调度任务,将其索引值赋给任务调度索引,将得分最大值赋给中间判断值;

S62,对所述得分最大的该类型调度任务进行再次调优,并记录调优后的耗时;

S63,将调优后的耗时与中间判断值进行比较,若调优后的耗时比中间判断值短,则加重对应的调度权重作为奖励,否则将减轻对应的调度权重作为惩罚。

4.根据权利要求3所述的在TVM中加速调优收敛的调度方法,其特征在于,所述奖励值大于惩罚值。

5.根据权利要求4所述的在TVM中加速调优收敛的调度方法,其特征在于,所述奖励内容为,在原调度权重的基础上加0.3;所述惩罚内容为,在原调度权重的基础上减0.2。

6.根据权利要求3所述的在TVM中加速调优收敛的调度方法,其特征在于,在所述S63中,还包括边界防护,当出现任一调度权重小于设定权重阈值时,进行调度任务进行重新洗牌。

7.根据权利要求6所述的在TVM中加速调优收敛的调度方法,其特征在于,所述重新洗牌的内容为:按照S3的方法,对当前的调度任务按照耗时长短为调度任务重新附上调度权重,耗时越长调度权重越重。

8.根据权利要求1所述的在TVM中加速调优收敛的调度方法,其特征在于,所述S1之前还包括:

S0,加载深度学习网络,并输入待检测图像,生成若干调度任务。

9.一种存储介质,其上存储有计算机程序,其特征在于,所述计算机程序被处理器执行时实现权利要求1至7中任一项所述的调度方法。

10.一种电子设备,其特征在于,包括:

处理器;以及

存储器,用于存储所述处理器的可执行指令;

其中,所述处理器配置为经由执行所述可执行指令来执行权利要求1至7中任一项所述的调度方法。

在TVM中加速调优收敛的调度方法、存储介质和电子设备

[0001] 相关申请的交叉引用

[0002] 本申请要求申请号为:2021106047694,申请日为:2021年5月31日,发明名称为:在TVM中加速调优收敛的调度方法、存储介质和电子设备的中国专利申请的优先权。

技术领域

[0003] 本发明涉及计算机技术领域,尤其涉及一种在TVM中加速调优收敛的调度方法、存储介质和电子设备。

背景技术

[0004] 在图像的深度学习网络中,通常包含多种任务,为了使整体网络性能更优,通常需要对任务进行调度。如文献:Zheng L,Jia C,Sun M,et al.Ansor:Generating High-Performance Tensor Programs for Deep Learning[J].2020即提出了两种策略用于调优任务的调度,分别是round-robin、gradient。其中Round-robin是类枚举调度方法,使用该方法每个任务都能够调度到,对于优化权重低和优化权重高的任务被调度的机会是一样的,这会导致非常耗时。gradient是基于梯度思想的调度策略,这种策略的收敛速度较快,但易陷入局部最优,不利于进一步的调度优化。

发明内容

[0005] 本发明的目的在于提供一种在TVM中加速调优收敛的调度方法,以解决背景技术中提到的至少一种技术问题。

[0006] 为实现上述目的,本发明提供如下技术方案:

[0007] 一种在TVM中加速调优收敛的调度方法,包括以下步骤:

[0008] S1,统计调度任务的类型,确定同类型调度任务的数量;

[0009] S2,首轮遍历调优,依次对调度任务进行首轮调优,记录调度任务对应的耗时;

[0010] S3,按照耗时长短为调度任务附上调度权重,耗时越长调度权重越重;

[0011] S4,计算耗时、同类型调度任务的数量、调度权重三者的乘积为同类型调度任务的得分;

[0012] S5,以所有调度任务的得分之和作为目标函数值与设定的目标阈值作比较,若目标函数值大于目标阈值,则跳至S6进行调度权重反馈,否则结束调优;

[0013] S6,调度权重反馈,对得分最大的同类型调度任务进行再次调优,并判断调优前后的耗时,若调优后的耗时比调优前的耗时短,则加重其调度权重作为奖励,否则减轻其调度权重作为惩罚;跳回至S4。

[0014] 进一步的,所述S2中,首轮遍历调优的具体内容为:初始化首轮遍历索引;对所述首轮遍历索引累加1,并依次对所述首轮遍历索引对应的调度任务进行调优直至所有调度任务完成首轮遍历调优。

[0015] 进一步的,所述S6中,调度权重反馈的具体内容为:

[0016] S61,选取得分最大的同类型调度任务,将其索引值赋给任务调度索引,将得分最大值赋给中间判断值;

[0017] S62,对所述得分最大的同类型调度任务进行再次调优,并记录调优后的耗时;

[0018] S63,将调优后的耗时与中间判断值进行比较,若调优后的耗时比中间判断值短,则加重对应的调度权重作为奖励,否则将减轻对应的调度权重作为惩罚。

[0019] 进一步的,所述奖励值大于惩罚值。

[0020] 进一步的,所述奖励内容为,在原调度权重的基础上加0.3;所述惩罚内容为,在原调度权重的基础上减0.2。

[0021] 进一步的,在所述S63中,还包括边界防护,当出现任一调度权重小于设定权重阈值时,进行调度任务进行重新洗牌。

[0022] 进一步的,所述重新洗牌的内容为:按照S3的方法,对当前的调度任务按照耗时长短为调度任务重新附上调度权重,耗时越长调度权重越重。

[0023] 进一步的,所述S1之前还包括:

[0024] S0,加载深度学习网络,并输入待检测图像,生成若干调度任务。

[0025] 一种存储介质,其上存储有计算机程序,所述计算机程序被处理器执行时实现上述的调度方法。

[0026] 一种电子设备,包括:

[0027] 处理器;以及

[0028] 存储器,用于存储所述处理器的可执行指令;

[0029] 其中,所述处理器配置为经由执行所述可执行指令来执行上述的调度方法。

[0030] 与现有技术相比,本发明的有益效果是:本发明引入调度权重反馈的设计,对调优后的耗时与调优前进行比较,并进行相应的调度权重奖励与惩罚,使的图像检测过程中,调度能够更加快速地收敛。另一方面,本发明还引入边界防护,在调度权重低于设定权重阈值的情况下,进行重新洗牌,防止陷入局部最优。

附图说明

[0031] 图1为本发明的方法流程图。

具体实施方式

[0032] 下面对本发明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发明一部分实施例,而不是全部的实施例。基于本发明中的实施例,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,都属于本发明保护的范围。

[0033] 请参照图1,本实施例提供一种在TVM中加速调优收敛的调度方法。

[0034] 调度目标函数及得分列表score[ ]设计:

[0035] (1)得分列表score[i]计算方式如下:

[0036] score[i]=task_costs[i]*task_wgts[i]*sort_wgts[i]

[0037] 其中:task_costs[i]是记录调度任务耗时的列表,task_wgts[i]是记录同类型调度任务数量的列表,sort_wgts[i]是记录调度权重的列表,i是列表的索引。

[0038] (2)调度目标函数F_{score}计算方式如下：

[0039] F_{score}＝sum(task_costs[i]*task_wgts[i]*sort_wgts[i])

[0040] 初始化操作。初始化首轮遍历调度索引idx＝0,调度权重反馈的中间判断值m_adj＝0,选择任务调度索引task_id＝-1,调度任务task对应的耗时列表task_costs[i]根据len(tasks)赋0初始化,调度权重列表sort_wgts[i]根据任务列表的长度len(tasks)赋0初始化;调度任务数量的列表Task_wgts[i]中记载的待调优的同类型子任务的个数,在深度学习网络定了之后也就固定了,跟调度调优过程无关。

[0041] 子任务的确定方式根据是否要对网络进行算子融合分为两种:(1)逐算子模式;(2)算子融合模式。

[0042] (1)逐算子模式：

[0043] 根据深度学习网络中包含的算子类型生成Task_wgts,此时每个算子为一个子任务,支持的算子包括卷积(conv2d)、池化(pool)、归一化(bn/in)、全连接(linear)、激活(relu)等常见算子。

[0044] (2)算子融合模式：

[0045] 支持卷积(conv2d)+归一化(bn)+激活(relu)融合成大算子(fused-conv2d-bn-relu),则此时的大算子(fused-conv2d-bn-relu)为一个独立子任务,其余未融合的算子为其他子任务。

[0046] 统计调度任务列表中属于同一类型(如卷积、池化、归一等等)的子任务数量,即构成同类型调度任务数量的列表Task_wgts[i]。

[0047] 具体包括以下步骤：

[0048] S0,加载深度学习网络,并输入待检测图像,生成若干调度任务;所述待检测图像可以应用于目标检测、人脸识别/检测、形体识别等。

[0049] S1,统计调度任务的类型,确定同类型调度任务的数量,构建同类型调度任务数量的列表Task_wgts[i];

[0050] S2,首轮遍历调优,依次对调度任务进行首轮调优,记录调度任务对应的耗时;具体为：

[0051] (1)当idx<len(tasks)时,对idx做+1操作,依次对任务进行调优,一轮下来得到填充满的记录每个任务task对应耗时的列表task_costs[i];

[0052] (2)当idx>＝len(tasks)时,跳出循环,对调度权重列表sort_wgts进行初始化;

[0053] S3,考虑到耗时更长的调度任务应该给予更大的调优机会,按照耗时长短为调度任务附上调度权重,耗时越长调度权重越重;

[0054] 如某一实施例中,涉及10种调度任务,其索引i为0-9,对应的耗时(单位/s)分别为0.2,0.5,0.3,0.4,1.1,0.6,2.3,0.9,1.4,1.8;

[0055] 预先设定调度权重1至10。

[0056] 按照耗时的长短排序,其调度权重列表sort_wgts为[1,4,2,3,7,5,10,6,8,9]。

[0057] S4,计算耗时、同类型调度任务的数量、调度权重三者的乘积为同类型调度任务的得分score[i];

[0058] score[i]＝task_costs[i]*task_wgts[i]*sort_wgts[i]

[0059] S5,以所有调度任务的得分之和作为目标函数值F_{score}与设定的目标阈值α作比较,

若目标函数值F_{score}大于目标阈值α,则跳至S6进行调度权重反馈,否则达到预期调优效果,跳出循环,结束调优;值得一提的是,目标阈值α为用户期望达到的网络推理效率,不同网络/不同硬件平台的设置数值不固定,由用户主观设定。

[0060] S6,调度权重反馈,对得分最大的同类型调度任务进行再次调优,并判断调优前后的耗时,若调优后的耗时比调优前的耗时短,则加重其调度权重作为奖励,否则减轻其调度权重作为惩罚;跳回至S4。

[0061] 调度权重反馈的具体内容为:

[0062] S61,选取得分score[i]最大的同类型调度任务,将其索引值i赋给任务调度索引task_id,将得分最大值score[task_id]赋给中间判断值m_adj;

[0063] S62,对所述得分最大的同类型调度任务进行再次调优,并记录调优后的耗时task_costs[task_id];

[0064] S63,将调优后的耗时task_costs[task_id]与中间判断值m_adj进行比较,若调优后的耗时比中间判断值短,说明在对这个task_id的调度任务进行调优后,获得了性能的提升,则加重对应的调度权重作为奖励,否则说明在对这个task_id的调度任务进行调优后,并没有获得性能的提升,将减轻对应的调度权重作为惩罚。另外考虑到调度的大部分轮次其实并不会对性能的提升有积极作用,所以奖励值略大于惩罚值,这样会让有希望性能提升的任务获得更多的调优机会。于一实施例中,所述奖励内容为,在原调度权重的基础上加0.3;所述惩罚内容为,在原调度权重的基础上减0.2。

[0065] 于一实施例中,为了避免调优陷入局部最优,还包括边界防护,当出现任一调度权重小于设定权重阈值时,进行调度任务进行重新洗牌,权重阈值优选为0.3。所述重新洗牌的内容为:按照S3的方法,对当前的调度任务按照耗时长短为调度任务重新附上调度权重[1-10],耗时越长调度权重越重,重新进行调优。

[0066] 通过以上的实施方式的描述,本领域的技术人员易于理解,这里描述的示例实施方式可以通过软件实现,也可以通过软件结合必要的硬件的方式来实现。因此,根据本公开实施方式的技术方案可以以软件产品的形式体现出来,该软件产品可以存储在一个非易失性存储介质(可以是CD-ROM,U盘,移动硬盘等)中或网络上,包括若干指令以使得一台电子设备(可以是个人计算机、服务器、终端装置、或者网络设备等)执行根据本公开实施方式的方法。

[0067] 对于本领域技术人员而言,显然本发明不限于上述示范性实施例的细节,而且在不背离本发明的精神或基本特征的情况下,能够以其他的具体形式实现本发明。因此,无论从哪一点来看,均应将实施例看作是示范性的,而且是非限制性的,本发明的范围由所附权利要求而不是上述说明限定,因此旨在将落在权利要求的等同要件的含义和范围内的所有变化囊括在本发明内。

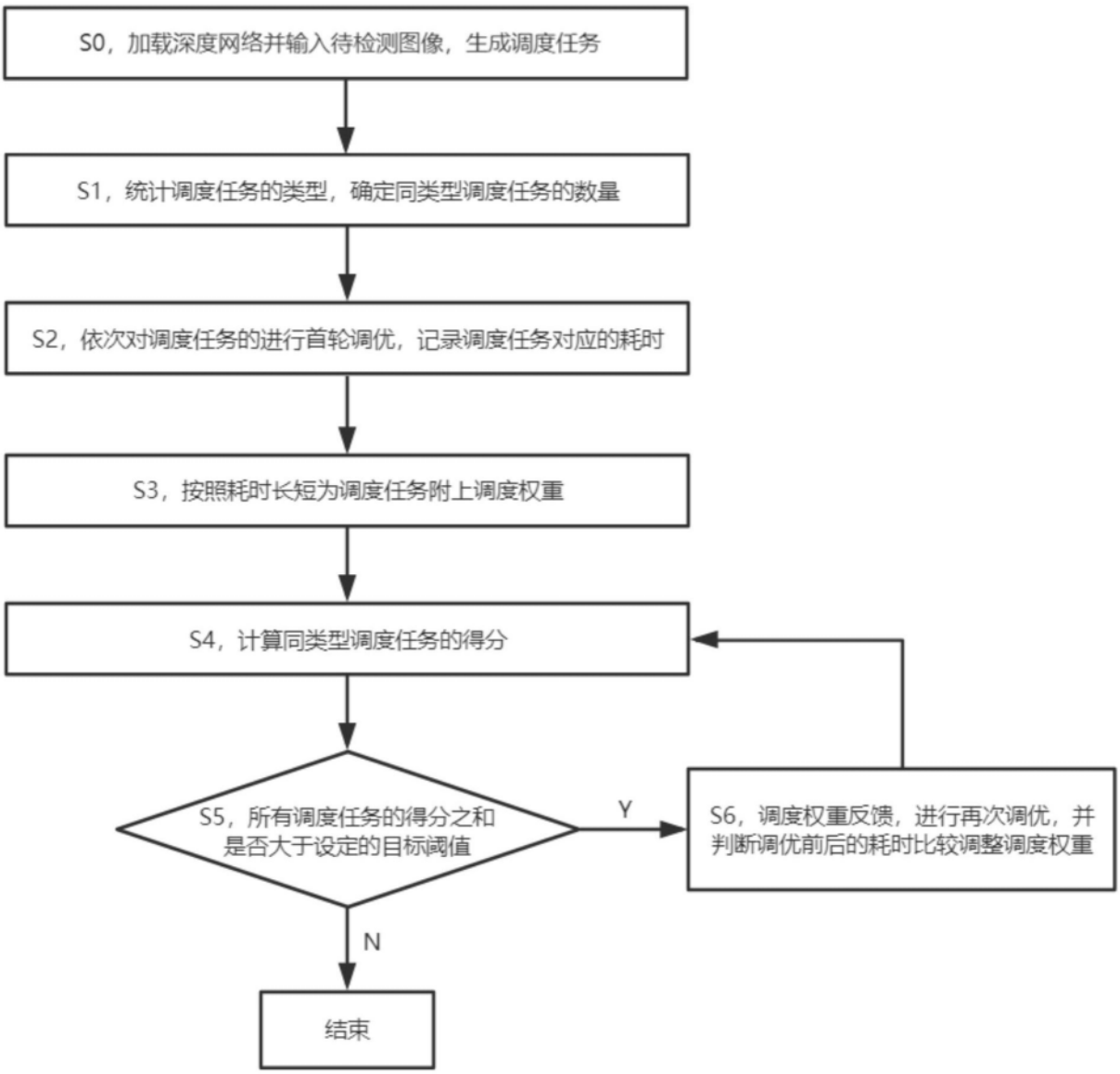

S0，加载深度网络并输入待检测图像，生成调度任务
S1，统计调度任务的类型，确定同类型调度任务的数量
S2，依次对调度任务的进行首轮调优，记录调度任务对应的耗时
S3，按照耗时长短为调度任务附上调度权重
S4，计算同类型调度任务的得分
S5，所有调度任务的得分之和
是否大于设定的目标阈值
Y
S6，调度权重反馈，进行再次调优，并
判断调优前后的耗时比较调整调度权重
N
结束

图1