(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 111240640 A
(43)申请公布日 2020.06.05

(21)申请号 202010071063.1

(22)申请日 2020.01.21

(71)申请人 苏州浪潮智能科技有限公司
地址 215100 江苏省苏州市吴中区吴中经济开发区郭巷街道官浦路1号9幢

(72)发明人 曹其春 赵雅倩 董刚 梁玲燕 尹文枫

(74)专利代理机构 北京集佳知识产权代理有限公司 11227
代理人 刘新雷

(51)Int.Cl.
G06F 8/10(2018.01)
G06N 20/00(2019.01)

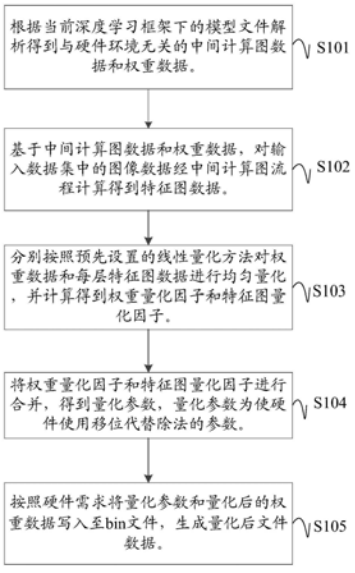
权利要求书2页 说明书9页 附图5页

(54)发明名称

基于硬件环境的数据量化方法、装置及可读存储介质

(57)摘要

本申请公开了一种基于硬件环境的数据量化方法、装置及计算机可读存储介质。其中，方法包括解析当前深度学习框架下的模型文件得到与硬件环境无关的中间计算图数据和权重数据，并对输入数据集中的图像数据经中间计算图流程计算得到特征图数据；分别按照预先设置的线性量化方法对权重数据和每层特征图数据进行均匀量化，计算得到权重量化因子和特征图量化因子，将权重量化因子和特征图量化因子进行合并，得到使硬件使用移位代替除法的量化参数；最后按照硬件需求将量化参数和量化后的权重数据写入至bin文件，生成量化后文件数据，从而解决了相关技术中为了支持多种深度学习框架导致量化软件包冗余、依赖库冲突的问题。



CN 111240640 A

1.一种基于硬件环境的数据量化方法,其特征在于,包括:

根据当前深度学习框架下的模型文件解析得到与硬件环境无关的中间计算图数据和权重数据;

基于所述中间计算图数据和所述权重数据,对输入数据集中的图像数据经中间计算图流程计算得到特征图数据;

分别按照预先设置的线性量化方法对所述权重数据和每层特征图数据进行均匀量化,并计算得到权重量化因子和特征图量化因子;

将所述权重量化因子和所述特征图量化因子进行合并,得到量化参数,所述量化参数为使硬件使用移位代替除法的参数;

按照硬件需求将所述量化参数和量化后的权重数据写入至bin文件,生成量化后文件数据。

2.根据权利要求1所述的基于硬件环境的数据量化方法,其特征在于,所述按照硬件需求将所述量化参数和量化后的权重数据写入至bin文件之前,还包括:

对所述量化参数和量化后的权重数据进行重排序,以使所述量化参数和量化后的权重数据的数据格式为64通道并行格式。

3.根据权利要求2所述的基于硬件环境的数据量化方法,其特征在于,所述解析得到当前深度学习框架的中间计算图数据和权重数据包括:

利用NNVM编译器中的NNVM组件解析所述模型文件得到所述中间计算图数据;

利用所述NNVM编译器中的TVM组件执行中间计算图的操作运算符并计算得到张量形式的权重数据。

4.根据权利要求3所述的基于硬件环境的数据量化方法,其特征在于,所述将所述权重量化因子和所述特征图量化因子进行合并为:

根据量化因子合并计算关系式将所述权重量化因子和所述特征图量化因子进行合并,所述量化因子合并计算关系式为:

$$y_w * y_f \approx \frac{1}{2^n};$$

式中,y_w为所述权重量化因子,y_f为所述特征图量化因子,n为所述量化参数。

5.根据权利要求1至4任意一项所述的基于硬件环境的数据量化方法,其特征在于,所述分别按照预先设置的线性量化方法对所述权重数据和每层特征图数据进行均匀量化,并计算得到权重量化因子和特征图量化因子包括:

计算每层特征图数据的平均值,以作为每层特征图平均数据;

统计所述权重数据和每层特征图平均数据的数据分布,并计算相应的限定值;

将所述权重数据和每层特征图平均数据限定在相应限定范围内,所述限定范围根据相应限定值确定;

将限定后的数据平均量化至int8数据精度的-127～+127之间,计算得到权重量化因子和特征图量化因子。

6.根据权利要求5所述的基于硬件环境的数据量化方法,其特征在于,所述计算相应的限定值包括:

所述权重数据的权重限定值根据权重限定值计算关系式计算得到,所述权重限定值计

算关系式为$x_w=\max(|w|)$，x_w为所述权重限定值，w为所述权重数据；相应的，所述权重数据的限定范围为$(-x_w,+x_w)$；

所述每层特征图平均数据的特征图限定值根据特征图限定值计算关系式计算得到，所述特征图限定值计算关系式为$x_f=\max(|F|)$，x_f为所述特征图限定值，F为每层特征图平均数据；相应的，每层特征图平均数据的限定范围为$(-x_f,+x_f)$。

7.一种基于硬件环境的数据量化装置，其特征在于，包括：

框架数据解析模块，用于根据当前深度学习框架下的模型文件解析得到与硬件环境无关的中间计算图数据和权重数据；

特征图数据计算模块，用于基于所述中间计算图数据和所述权重数据，对输入数据集中的图像数据经中间计算图流程计算得到特征图数据；

线性量化模块，用于分别按照预先设置的线性量化方法对所述权重数据和每层特征图数据进行均匀量化，并计算得到权重量化因子和特征图量化因子；

量化参数计算模块，用于将所述权重量化因子和所述特征图量化因子进行合并，得到量化参数，所述量化参数为使硬件使用移位代替除法的参数；

硬件可识别数据输出模块，用于按照硬件需求将所述量化参数和量化后的权重数据写入至bin文件，生成量化后文件数据。

8.根据权利要求7所述的基于硬件环境的数据量化装置，其特征在于，还包括重排序模块，所述重排序模块用于对所述量化参数和量化后的权重数据进行重排序，以使所述量化参数和量化后的权重数据的数据格式为64通道并行格式。

9.一种基于硬件环境的数据量化装置，其特征在于，包括处理器，所述处理器用于执行存储器中存储的计算机程序时实现如权利要求1至6任一项所述基于硬件环境的数据量化方法的步骤。

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上存储有基于硬件环境的数据量化程序，所述基于硬件环境的数据量化程序被处理器执行时实现如权利要求1至6任一项所述基于硬件环境的数据量化方法的步骤。

基于硬件环境的数据量化方法、装置及可读存储介质

技术领域

[0001] 本申请涉及人工智能技术领域，特别是涉及一种基于硬件环境的数据量化方法、装置及计算机可读存储介质。

背景技术

[0002] 随着人工智能在各个领域的发展，如农业、金融、安防、健康医疗、制造等，用户对基于人工智能技术的产品的计算速度、精度和功耗有更高的需求。各大硬件产商研发专门针对人工智能算法计算的加速卡及相应配套的量化方案，来加速人工智能算法在日常使用的普及。

[0003] AI(Artificial Intelligence，人工智能)加速卡的大规模和并行特点，导致AI加速卡的开发也极具挑战，同时还需要满足量化方案能够使用低精度运算实现类似高精度运算的算法精度。为了满足高精度数据映射到低精度、减少硬件资源开销，需提前对高精度数据进行量化以生成低精度的权重数据和量化参数文件，软件端的量化工具包的开发即满足上述需求。

[0004] 但是，随着目前深度学习框架种类增多，为AI加速卡适应各种框架下的模型增加了困难，普通的量化工具包为兼容各种框架，需提前安装多种深度学习框架软件，很容易造成主机端软件的冗余和各种依赖库的冲突。

[0005] 鉴于此，如何解决为了支持多种深度学习框架导致软件包冗余、依赖库冲突的问题，是本领域技术人员需要解决的技术问题。

发明内容

[0006] 本申请提供了一种基于硬件环境的数据量化方法、装置及计算机可读存储介质，解决了相关技术中为了支持多种深度学习框架导致软件包冗余、依赖库冲突的问题。

[0007] 为解决上述技术问题，本发明实施例提供以下技术方案：

[0008] 本发明实施例一方面提供了一种基于硬件环境的数据量化方法，包括：

[0009] 根据当前深度学习框架下的模型文件解析得到与硬件环境无关的中间计算图数据和权重数据；

[0010] 基于所述中间计算图数据和所述权重数据，对输入数据集中的图像数据经中间计算图流程计算得到特征图数据；

[0011] 分别按照预先设置的线性量化方法对所述权重数据和每层特征图数据进行均匀量化，并计算得到权重量化因子和特征图量化因子；

[0012] 将所述权重量化因子和所述特征图量化因子进行合并，得到量化参数，所述量化参数为使硬件使用移位代替除法的参数；

[0013] 按照硬件需求将所述量化参数和量化后的权重数据写入至bin文件，生成量化后文件数据。

[0014] 可选的，所述按照硬件需求将所述量化参数和量化后的权重数据写入至bin文件

之前，还包括：

[0015] 对所述量化参数和量化后的权重数据进行重排序，以使所述量化参数和量化后的权重数据的数据格式为64通道并行格式。

[0016] 可选的，所述解析得到当前深度学习框架的中间计算图数据和权重数据包括：

[0017] 利用NNVM编译器中的NNVM组件解析所述模型文件得到所述中间计算图数据；

[0018] 利用所述NNVM编译器中的TVM组件执行中间计算图的操作运算符并计算得到张量形式的权重数据。

[0019] 可选的，所述将所述权重量化因子和所述特征图量化因子进行合并为：

[0020] 根据量化因子合并计算关系式将所述权重量化因子和所述特征图量化因子进行合并，所述量化因子合并计算关系式为：

[0021] $$y_w * y_f \approx \frac{1}{2^n};$$

[0022] 式中，y_w为所述权重量化因子，y_f为所述特征图量化因子，n为所述量化参数。

[0023] 可选的，所述分别按照预先设置的线性量化方法对所述权重数据和每层特征图数据进行均匀量化，并计算得到权重量化因子和特征图量化因子包括：

[0024] 计算每层特征图数据的平均值，以作为每层特征图平均数据；

[0025] 统计所述权重数据和每层特征图平均数据的数据分布，并计算相应的限定值；

[0026] 将所述权重数据和每层特征图平均数据限定在相应限定范围内，所述限定范围根据相应限定值确定；

[0027] 将限定后的数据平均量化至int8数据精度的-127～+127之间，计算得到权重量化因子和特征图量化因子。

[0028] 可选的，所述计算相应的限定值包括：

[0029] 所述权重数据的权重限定值根据权重限定值计算关系式计算得到，所述权重限定值计算关系式为$x_w=\max(|w|)$，x_w为所述权重限定值，w为所述权重数据；相应的，所述权重数据的限定范围为$(-x_w,+x_w)$；

[0030] 所述每层特征图平均数据的特征图限定值根据特征图限定值计算关系式计算得到，所述特征图限定值计算关系式为$x_f=\max(|F|)$，x_f为所述特征图限定值，F为每层特征图平均数据；相应的，每层特征图平均数据的限定范围为$(-x_f,+x_f)$。

[0031] 本发明实施例另一方面提供了一种基于硬件环境的数据量化装置，包括：

[0032] 框架数据解析模块，用于根据当前深度学习框架下的模型文件解析得到与硬件环境无关的中间计算图数据和权重数据；

[0033] 特征图数据计算模块，用于基于所述中间计算图数据和所述权重数据，对输入数据集中的图像数据经中间计算图流程计算得到特征图数据；

[0034] 线性量化模块，用于分别按照预先设置的线性量化方法对所述权重数据和每层特征图数据进行均匀量化，并计算得到权重量化因子和特征图量化因子；

[0035] 量化参数计算模块，用于将所述权重量化因子和所述特征图量化因子进行合并，得到量化参数，所述量化参数为使硬件使用移位代替除法的参数；

[0036] 硬件可识别数据输出模块，用于按照硬件需求将所述量化参数和量化后的权重数据写入至bin文件，生成量化后文件数据。

[0037] 可选的,还包括重排序模块,所述重排序模块用于对所述量化参数和量化后的权重数据进行重排序,以使所述量化参数和量化后的权重数据的数据格式为64通道并行格式。

[0038] 本发明实施例还提供了一种基于硬件环境的数据量化装置,包括处理器,所述处理器用于执行存储器中存储的计算机程序时实现如前任一项所述基于硬件环境的数据量化方法的步骤。

[0039] 本发明实施例最后还提供了一种计算机可读存储介质,所述计算机可读存储介质上存储有基于硬件环境的数据量化程序,所述基于硬件环境的数据量化程序被处理器执行时实现如前任一项所述基于硬件环境的数据量化方法的步骤。

[0040] 本申请提供的技术方案的优点在于,将深度学习框架下的模型文件转化为与硬件无关的中间计算图数据和权重数据,从而可支持各种深度学习框架在不同的计算机平台上运行;采用线性量化策略,对每层特征图数据和权重数据进行均匀量化,保持最少的量化参数,同时合并量化参数利于硬件推理,数据都写入硬件能够识别的bin文件,从而解决了相关技术由于支持多种深度学习框架带来的软件冗余、依赖库冲突问题,可有效减少为支持多种深度学习框架而开发的各种接口,精简了主机端软件的工作量和开发难度;还可减少硬件计算资源,加速AI加速卡推理速度,降低能耗。

[0041] 此外,本发明实施例还针对基于硬件环境的数据量化方法提供了相应的实现装置及计算机可读存储介质,进一步使得所述方法更具有实用性,所述装置及计算机可读存储介质具有相应的优点。

[0042] 应当理解的是,以上的一般描述和后文的细节描述仅是示例性的,并不能限制本公开。

附图说明

[0043] 为了更清楚的说明本发明实施例或相关技术的技术方案,下面将对实施例或相关技术描述中所需要使用的附图作简单的介绍,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。

[0044] 图1为本发明实施例提供的一种基于硬件环境的数据量化方法的流程示意图;

[0045] 图2为本发明实施例提供的另一种基于硬件环境的数据量化方法的流程示意图;

[0046] 图3为本发明实施例提供的重排序后的数据显示示意图;

[0047] 图4为本发明实施例提供的基于硬件环境的数据量化装置的一种具体实施方式结构图;

[0048] 图5为本发明实施例提供的基于硬件环境的数据量化装置的另一种具体实施方式结构图;

[0049] 图6为本发明实施例提供的基于硬件环境的数据量化装置的再一种具体实施方式结构图。

具体实施方式

[0050] 为了使本技术领域的人员更好地理解本发明方案,下面结合附图和具体实施方式

对本发明作进一步的详细说明。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。

[0051] 本申请的说明书和权利要求书及上述附图中的术语“第一”、“第二”、“第三”“第四”等是用于区别不同的对象，而不是用于描述特定的顺序。此外术语“包括”和“具有”以及他们任何变形，意图在于覆盖不排他的包含。例如包含了一系列步骤或单元的过程、方法、系统、产品或设备没有限定于已列出的步骤或单元，而是可包括没有列出的步骤或单元。

[0052] 在介绍了本发明实施例的技术方案后，下面详细的说明本申请的各种非限制性实施方式。

[0053] 首先参见图1，图1为本发明实施例提供的一种基于硬件环境的数据量化方法的流程示意图，本发明实施例可包括以下内容：

[0054] S101：根据当前深度学习框架下的模型文件解析得到与硬件环境无关的中间计算图数据和权重数据。

[0055] 在本申请中，深度学习框架可为任何一种目前现有的深度学习框架，加载深度学习框架下的模型文件如tensorflow框架下的pb文件。可利用任何一种现有方法解析得到中间计算图和权重数据，这均不影响本申请的实现，例如可利用NNVM编译器的NNVM组件将不同框架模型文件转成框架无关的计算中间图，再使用TVM组件执行中间图的操作运算符，解除计算图的各种操作运算与硬件的无关性，使得本申请可支持各种深度学习框架和在不同的计算机平台上运行。NNVM编译器包括基于TVM堆栈中的两个组件，处理中间计算图的NNVM组件和处理张量操作运算符的TVM组件。NNVM组件(计算图媒介表示堆栈)可用于将来自不同框架的工作命令表示为标准化的计算图，然后将这些高级计算图转换为执行图。将中间计算图以一种与框架无关的形式表达出来的想法。TVM组件(张量媒介表示堆栈)的执行对象是计算图中的操作运算符，它把操作运算符优化成对应目标后端硬件的操作运算符。它与NNVM组件不同，提供了一种与硬件无关的、对应特定领域的语言，以简化在张量索引级别中的操作符执行。

[0056] S102：基于中间计算图数据和权重数据，对输入数据集中的图像数据经中间计算图流程计算得到特征图数据。

[0057] 其中，输入数据集可为S101相应深度学习框架下的训练数据集，输入数据集包含的图像总数本申请不作任何限定，例如可为包含2000张图像的数据集。在获取得到输入数据集，为了便于后续图像处理，还可对输入数据集中的图像数据进行图像预处理，图像预处理例如可先进行图层处理，然后将图像数据均统一转化为float类型数据，最后还可进行平移处理，平移值可为0-255之间的任何一个值。在TVM框架的基础操作下，可对输入图像数据算出计算图每层的输出数据，也即得到每层特征图数据，可将计算得到的每层特征图数据保存在内存中，累加计算结果，然后计算每特征图数据的平均值。

[0058] S103：分别按照预先设置的线性量化方法对权重数据和每层特征图数据进行均匀量化，并计算得到权重量化因子和特征图量化因子。

[0059] 在本申请中，可采用任何一种线性量化方法对数据进行量化处理，本申请对此不做任何限定。举例来说，对于AI加速卡使用int8数据精度代替float数据精度，采用针对每层数据的线性量化方法，统计每层特征图数据和权重数据分布，将数据限定在-X～+X之间，

再平均量化到int8的-127～+127之间，在硬件计算推理过程中量化参数合到一个，并近似为硬件能够使用移位代替除法的参数。

[0060] 其中，权重量化因子和特征图量化因子根据相应的线性量化方法和原始数据计算得到，此处的，原始数据为指权重数据或每层特征图数据。

[0061] S104：将权重量化因子和特征图量化因子进行合并，得到量化参数，量化参数为使硬件使用移位代替除法的参数。

[0062] 在本发明实施例中，在S103根据线性量化方法统计计算图每层输出数据和权重数据分布，计算得到合理的量化参数，最终得到的量化参数可使硬件在进行推理时使用移位代替除法，量化参数例如可为近似为2的倍数，也即量化参数作为硬件推理的移位参数。可为应用于任何一种硬件中，例如FPGA。

[0063] S105：按照硬件需求将量化参数和量化后的权重数据写入至bin文件，生成量化后文件数据。

[0064] 可以理解的是，本申请基于硬件环境，为了实现硬件可以识别这些数据，并在数据推理时使用，可将按照硬件需求将量化参数和量化后的权重数据写入至硬件能够识别的bin文件中。

[0065] 在本发明实施例提供的技术方案中，将深度学习框架下的模型文件转化为与硬件无关的中间计算图数据和权重数据，从而可支持各种深度学习框架在不同的计算机平台上运行；采用线性量化策略，对每层特征图数据和权重数据进行均匀量化，保持最少的量化参数，同时合并量化参数利于硬件推理，数据都写入硬件能够识别的bin文件，从而解决了相关技术由于支持多种深度学习框架带来的软件冗余、依赖库冲突问题，可有效减少为支持多种深度学习框架而开发的各种接口，精简了主机端软件的工作量和开发难度；还可减少硬件计算资源，加速AI加速卡推理速度，降低能耗。

[0066] 此外，本申请还提供了另外一个实施例，请参见图2，图2为本发明实施例提供的另一种基于硬件环境的数据量化方法的流程示意图，本发明实施例例如可应用于基于FPGA(Field-Programmable Gate Array现场可编程门阵列)的AI加速卡在int8数据精度的量化，具体的可包括以下内容：

[0067] S201：利用NNVM编译器解析当前深度学习框架下的模型文件得到与硬件环境无关的中间计算图数据和权重数据。

[0068] 在该步骤中，可利用NNVM编译器中的NNVM组件解析模型文件得到中间计算图数据；利用NNVM编译器中的TVM组件执行中间计算图的操作运算符并计算得到张量形式的权重数据，从而得到与硬件毫无关系的数据，不用受限于所使用的硬件环境。

[0069] S202：基于中间计算图数据和权重数据，对输入数据集中的图像数据经中间计算图流程计算得到特征图数据，计算每层特征图数据的平均值，以作为每层特征图平均数据。

[0070] S203：统计权重数据和每层特征图平均数据的数据分布，并计算相应的限定值。

[0071] 具体地，权重数据的权重限定值可根据权重限定值计算关系式计算得到，权重限定值计算关系式为$x_w=\max(|w|)$，x_w为权重限定值，w为权重数据。每层特征图平均数据的特征图限定值可根据特征图限定值计算关系式计算得到，特征图限定值计算关系式为$x_f=\max(|F|)$，x_f为特征图限定值，F为每层特征图平均数据。

[0072] S204：将权重数据和每层特征图平均数据限定在相应限定范围内。

[0073] 本发明实施例的限定范围根据相应限定值确定，基于S203计算得到的限定值，权重数据的限定范围可为(-x_w,+x_w)；每层特征图平均数据的限定范围可为(-x_f,+x_f)。

[0074] S205：将限定后的数据平均量化至int8数据精度的-127～+127之间，计算得到权重量化因子和特征图量化因子。

[0075] 在经过S203和S204之后，$\frac{x_w^{'}}{y_w}=127$，$y_w=x_w^{'}/127$；$\frac{x_f^{'}}{y_f}=127$，$y_f=x'_f/127$；x'_w、x'_f为量化后的权重数据和特征图数据，y_w为权重量化因子，y_f为特征图量化因子。

[0076] S206：根据量化因子合并计算关系式将权重量化因子和特征图量化因子进行合并，量化因子合并计算关系式为：

[0077] $$y_w * y_f \approx \frac{1}{2^n};$$

[0078] 式中，y_w为权重量化因子，y_f为特征图量化因子，n为量化参数。

[0079] S207：对量化参数和量化后的权重数据进行重排序，以使量化参数和量化后的权重数据的数据格式为64通道并行格式。

[0080] 使用FPGA开发的AI加速卡，为使硬件资源的最大化利用，便于硬件64通道并行的计算操作，量化参数和量化后的权重数据满足硬件64通道并行的策略，可对数据进行重排序，生成如图3所示二进制的bin文件。如此在将数据输入到硬件进行推理时，无需转换数据格式，即可将数据平均分配到64并行通道中进行计算，减少硬件在数据转换上的资源使用。

[0081] S208：按照硬件需求将重排序的量化参数和量化后的权重数据写入至bin文件，生成量化后文件数据。

[0082] 本发明实施例与上述发明实施例相应的实施方法和相同的实现步骤可参阅上述实施例的描述，此处，便在赘述。

[0083] 由上可知，本发明实施解决了相关技术中为了支持多种深度学习框架导致软件包冗余、依赖库冲突的问题，可有效精简主机端软件的工作量和开发难度，减少硬件计算资源，加速AI加速卡推理速度，降低能耗。

[0084] 需要说明的是，本申请中各步骤之间没有严格的先后执行顺序，只要符合逻辑上的顺序，则这些步骤可以同时执行，也可按照某种预设顺序执行，图1-图2只是一种示意方式，并不代表只能是这样的执行顺序。

[0085] 本发明实施例还针对基于硬件环境的数据量化方法提供了相应的装置，进一步使得所述方法更具有实用性。其中，装置可从功能模块的角度和硬件的角度分别说明。下面对本发明实施例提供的基于硬件环境的数据量化装置进行介绍，下文描述的基于硬件环境的数据量化装置与上文描述的基于硬件环境的数据量化方法可相互对应参照。

[0086] 基于功能模块的角度，参见图4，图4为本发明实施例提供的基于硬件环境的数据量化装置在一种具体实施方式下的结构图，该装置可包括：

[0087] 框架数据解析模块401，用于根据当前深度学习框架下的模型文件解析得到与硬件环境无关的中间计算图数据和权重数据。

[0088] 特征图数据计算模块402，用于基于中间计算图数据和权重数据，对输入数据集中的图像数据经中间计算图流程计算得到特征图数据。

[0089] 线性量化模块403,用于分别按照预先设置的线性量化方法对权重数据和每层特征图数据进行均匀量化,并计算得到权重量化因子和特征图量化因子。

[0090] 量化参数计算模块404,用于将权重量化因子和特征图量化因子进行合并,得到量化参数,量化参数为使硬件使用移位代替除法的参数。

[0091] 硬件可识别数据输出模块405,用于按照硬件需求将量化参数和量化后的权重数据写入至bin文件,生成量化后文件数据。

[0092] 可选的,在本实施例的一些实施方式中,请参阅图5,所述装置例如还可以包括重排序模块406,用于对量化参数和量化后的权重数据进行重排序,以使量化参数和量化后的权重数据的数据格式为64通道并行格式。

[0093] 在本实施例的另一些实施方式中,所述框架数据解析模块401可具体用于利用NNVM编译器中的NNVM组件解析模型文件得到中间计算图数据;利用NNVM编译器中的TVM组件执行中间计算图的操作运算符并计算得到张量形式的权重数据。

[0094] 在本实施例的其他一些实施方式中,所述线性量化模块403可包括:

[0095] 平均值计算子模块,用于计算每层特征图数据的平均值,以作为每层特征图平均数据;

[0096] 限定值计算子模块,用于统计权重数据和每层特征图平均数据的数据分布,并计算相应的限定值;

[0097] 数据限定子模块,用于将权重数据和每层特征图平均数据限定在相应限定范围内,限定范围根据相应限定值确定;

[0098] 量化子模块,用于将限定后的数据平均量化至int8数据精度的-127～+127之间,计算得到权重量化因子和特征图量化因子。

[0099] 本发明实施例所述基于硬件环境的数据量化装置的各功能模块的功能可根据上述方法实施例中的方法具体实现,其具体实现过程可以参照上述方法实施例的相关描述,此处不再赘述。

[0100] 由上可知,本发明实施解决了相关技术中为了支持多种深度学习框架导致软件包冗余、依赖库冲突的问题,可有效精简主机端软件的工作量和开发难度,减少硬件计算资源,加速AI加速卡推理速度,降低能耗。

[0101] 上文中提到的基于硬件环境的数据量化装置是从功能模块的角度描述,进一步的,本申请还提供一种基于硬件环境的数据量化装置,是从硬件角度描述。图6为本申请实施例提供的另一种基于硬件环境的数据量化装置的结构图。如图6所示,该装置包括存储器60,用于存储计算机程序;

[0102] 处理器61,用于执行计算机程序时实现如上述任一实施例提到的基于硬件环境的数据量化方法的步骤,其中计算机程序例如可利用python语言来编译实现。

[0103] 其中,处理器61可以包括一个或多个处理核心,比如4核心处理器、8核心处理器等。处理器61可以采用DSP(Digital Signal Processing,数字信号处理)、FPGA(Field－Programmable Gate Array,现场可编程门阵列)、PLA(Programmable Logic Array,可编程逻辑阵列)中的至少一种硬件形式来实现。处理器61也可以包括主处理器和协处理器,主处理器是用于对在唤醒状态下的数据进行处理的处理器,也称CPU(Central Processing Unit,中央处理器);协处理器是用于对在待机状态下的数据进行处理的低功耗处理器。在

一些实施例中，处理器61可以在集成有GPU(Graphics Processing Unit，图像处理器)，GPU用于负责显示屏所需要显示的内容的渲染和绘制。一些实施例中，处理器61还可以包括AI(Artificial Intelligence，人工智能)处理器，该AI处理器用于处理有关机器学习的计算操作。

[0104] 存储器60可以包括一个或多个计算机可读存储介质，该计算机可读存储介质可以是非暂态的。存储器60还可包括高速随机存取存储器，以及非易失性存储器，比如一个或多个磁盘存储设备、闪存存储设备。本实施例中，存储器60至少用于存储以下计算机程序601，其中，该计算机程序被处理器61加载并执行之后，能够实现前述任一实施例公开的基于硬件环境的数据量化方法的相关步骤。另外，存储器60所存储的资源还可以包括操作系统602和数据603等，存储方式可以是短暂存储或者永久存储。其中，操作系统602可以包括Windows、Unix、Linux等。数据603可以包括但不限于测试结果对应的数据等。

[0105] 在一些实施例中，基于硬件环境的数据量化装置还可包括有显示屏62、输入输出接口63、通信接口64、电源65以及通信总线66。

[0106] 本领域技术人员可以理解，图6中示出的结构并不构成对基于硬件环境的数据量化装置的限定，可以包括比图示更多或更少的组件，例如传感器67。

[0107] 本发明实施例所述基于硬件环境的数据量化装置的各功能模块的功能可根据上述方法实施例中的方法具体实现，其具体实现过程可以参照上述方法实施例的相关描述，此处不再赘述。

[0108] 由上可知，本发明实施解决了相关技术中为了支持多种深度学习框架导致软件包冗余、依赖库冲突的问题，可有效精简主机端软件的工作量和开发难度，减少硬件计算资源，加速AI加速卡推理速度，降低能耗。

[0109] 可以理解的是，如果上述实施例中的基于硬件环境的数据量化方法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，本申请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，执行本申请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)、电可擦除可编程ROM、寄存器、硬盘、可移动磁盘、CD-ROM、磁碟或者光盘等各种可以存储程序代码的介质。

[0110] 基于此，本发明实施例还提供了一种计算机可读存储介质，存储有基于硬件环境的数据量化程序，所述基于硬件环境的数据量化程序被处理器执行时如上任意一实施例所述基于硬件环境的数据量化方法的步骤。

[0111] 本发明实施例所述计算机可读存储介质的各功能模块的功能可根据上述方法实施例中的方法具体实现，其具体实现过程可以参照上述方法实施例的相关描述，此处不再赘述。

[0112] 由上可知，本发明实施解决了相关技术中为了支持多种深度学习框架导致软件包冗余、依赖库冲突的问题，可有效精简主机端软件的工作量和开发难度，减少硬件计算资源，加速AI加速卡推理速度，降低能耗。

[0113] 本说明书中各个实施例采用递进的方式描述，每个实施例重点说明的都是与其它

实施例的不同之处，各个实施例之间相同或相似部分互相参见即可。对于实施例公开的装置而言，由于其与实施例公开的方法相对应，所以描述的比较简单，相关之处参见方法部分说明即可。

[0114] 专业人员还可以进一步意识到，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、计算机软件或者二者的结合来实现，为了清楚地说明硬件和软件的可互换性，在上述说明中已经按照功能一般性地描述了各示例的组成及步骤。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。

[0115] 以上对本申请所提供的一种基于硬件环境的数据量化方法、装置及计算机可读存储介质进行了详细介绍。本文中应用了具体个例对本发明的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本发明的方法及其核心思想。应当指出，对于本技术领域的普通技术人员来说，在不脱离本发明原理的前提下，还可以对本申请进行若干改进和修饰，这些改进和修饰也落入本申请权利要求的保护范围内。

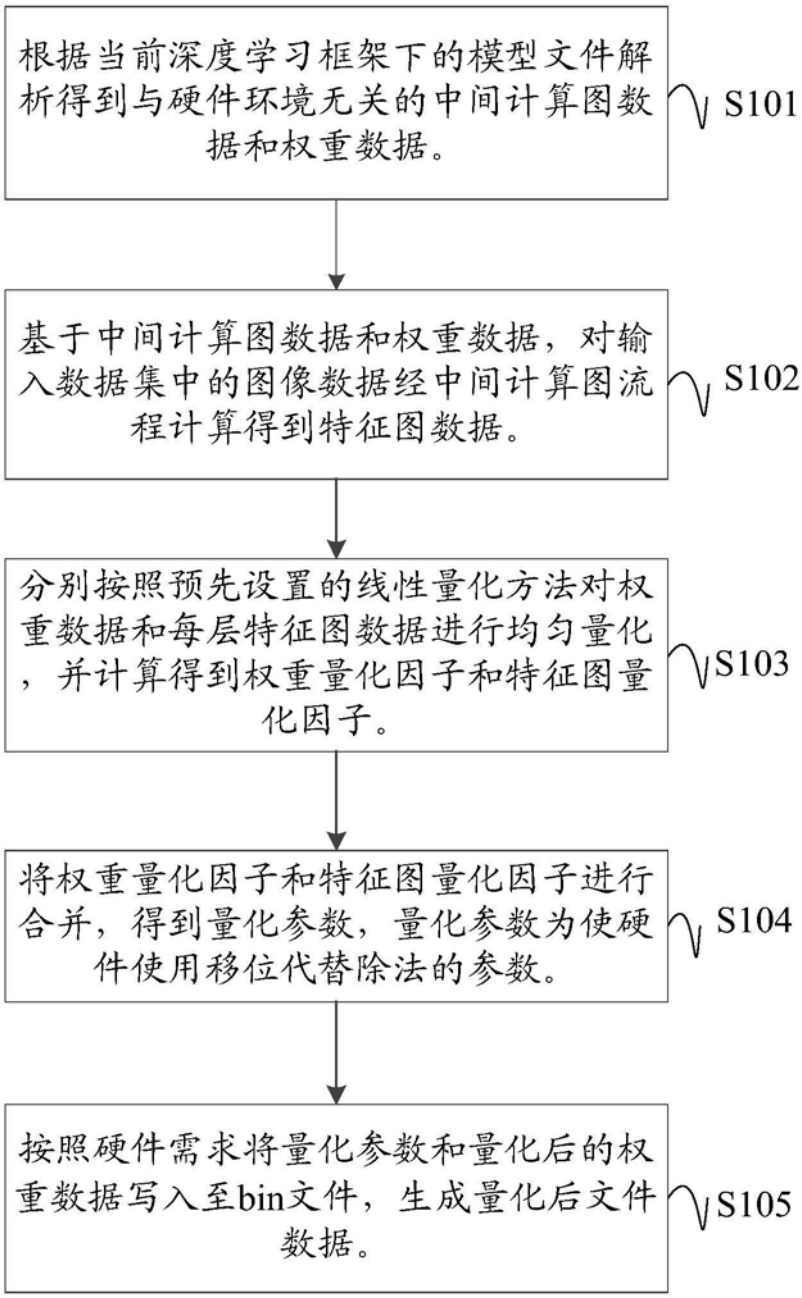

根据当前深度学习框架下的模型文件解析得到与硬件环境无关的中间计算图数据和权重数据。
S101
基于中间计算图数据和权重数据，对输入数据集中的图像数据经中间计算图流程计算得到特征图数据。
S102
分别按照预先设置的线性量化方法对权重数据和每层特征图数据进行均匀量化，并计算得到权重量化因子和特征图量化因子。
S103
将权重量化因子和特征图量化因子进行合并，得到量化参数，量化参数为使硬件使用移位代替除法的参数。
S104
按照硬件需求将量化参数和量化后的权重数据写入至bin文件，生成量化后文件数据。
S105

图1

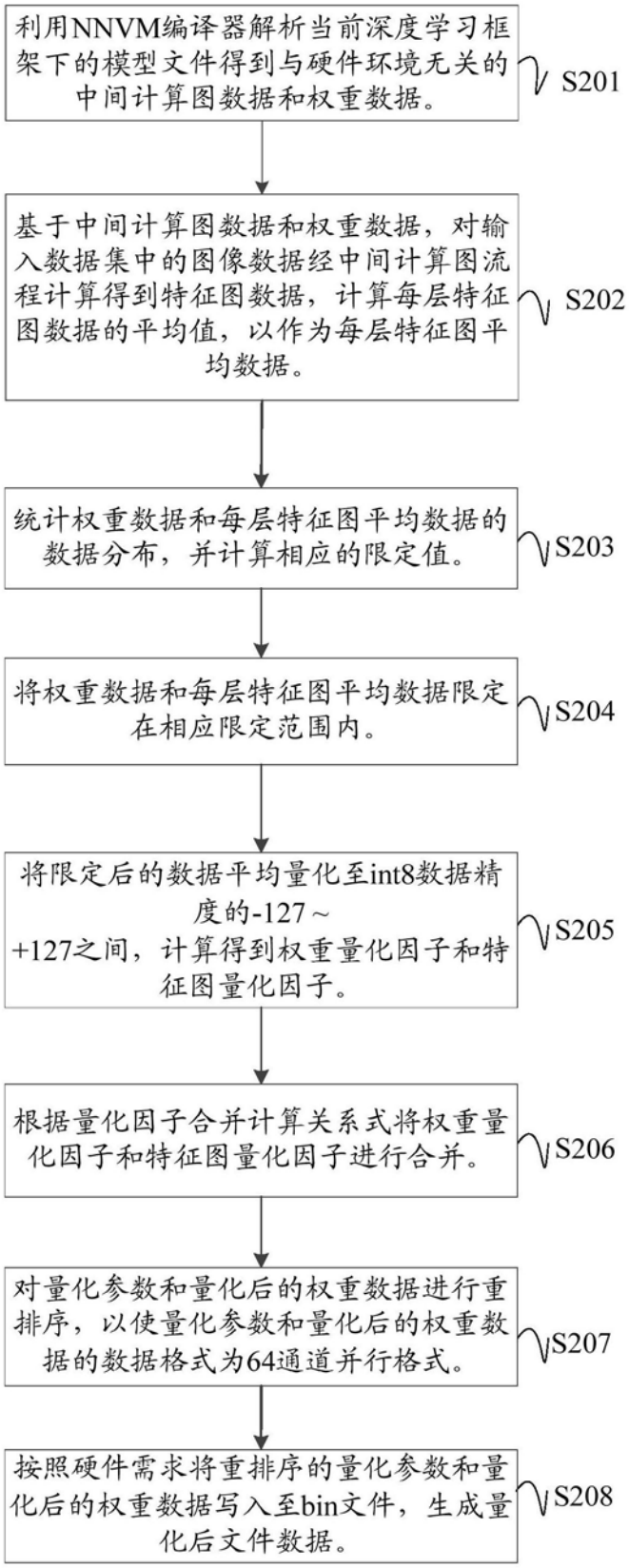

利用NNVM编译器解析当前深度学习框架下的模型文件得到与硬件环境无关的中间计算图数据和权重数据。
S201
基于中间计算图数据和权重数据，对输入数据集中的图像数据经中间计算图流程计算得到特征图数据，计算每层特征图数据的平均值，以作为每层特征图平均数据。
S202
统计权重数据和每层特征图平均数据的数据分布，并计算相应的限定值。
S203
将权重数据和每层特征图平均数据限定在相应限定范围内。
S204
将限定后的数据平均量化至int8数据精度的-127～+127之间，计算得到权重量化因子和特征图量化因子。
S205
根据量化因子合并计算关系式将权重量化因子和特征图量化因子进行合并。
S206
对量化参数和量化后的权重数据进行重排序，以使量化参数和量化后的权重数据的数据格式为64通道并行格式。
S207
按照硬件需求将重排序的量化参数和量化后的权重数据写入至bin文件，生成量化后文件数据。
S208

图2

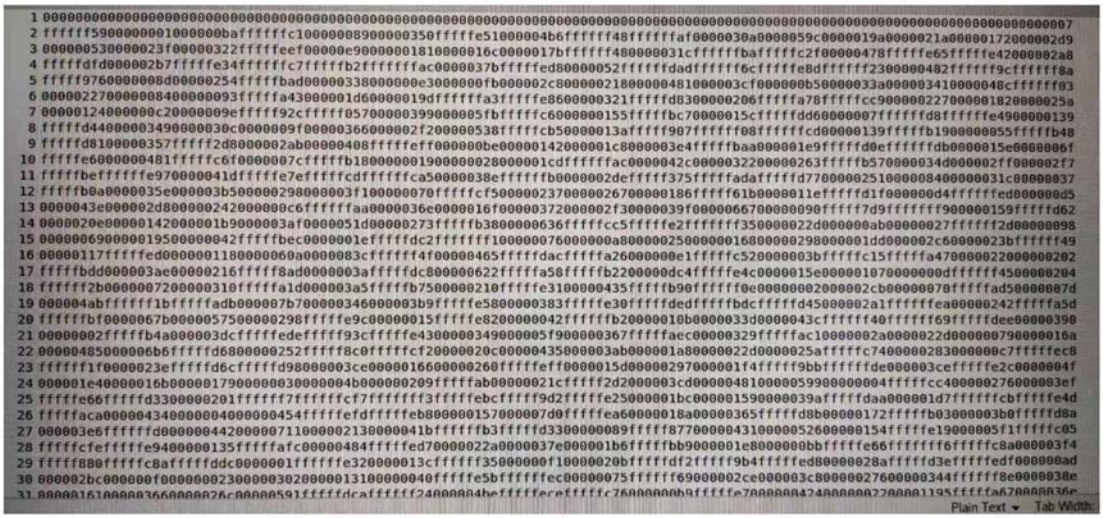


图3

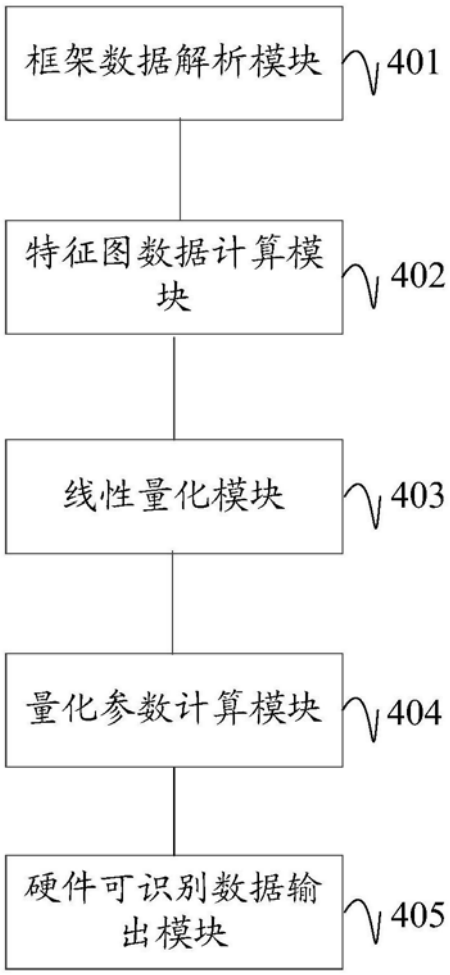

框架数据解析模块
401
特征图数据计算模块
402
线性量化模块
403
量化参数计算模块
404
硬件可识别数据输出模块
405

图4

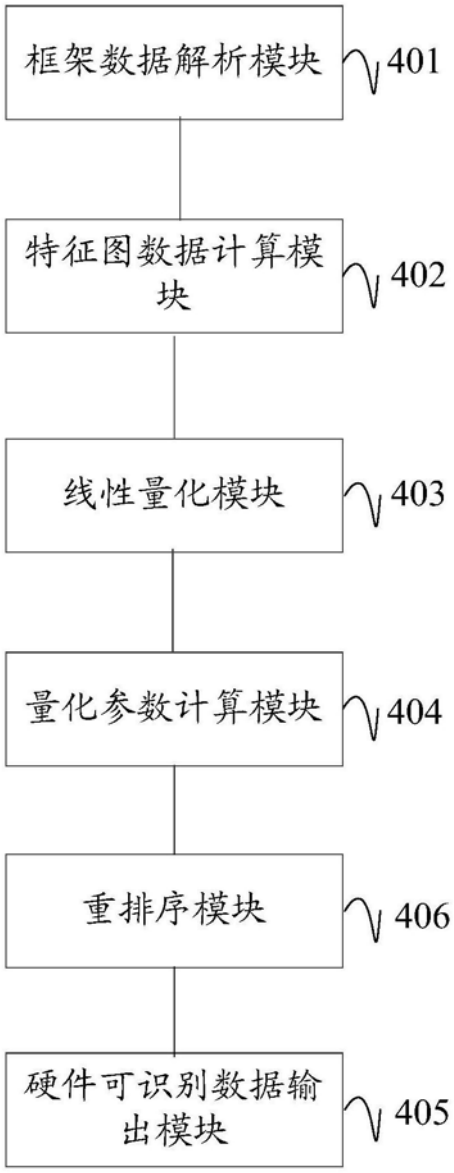

框架数据解析模块
401
特征图数据计算模块
402
线性量化模块
403
量化参数计算模块
404
重排序模块
406
硬件可识别数据输出模块
405

图5

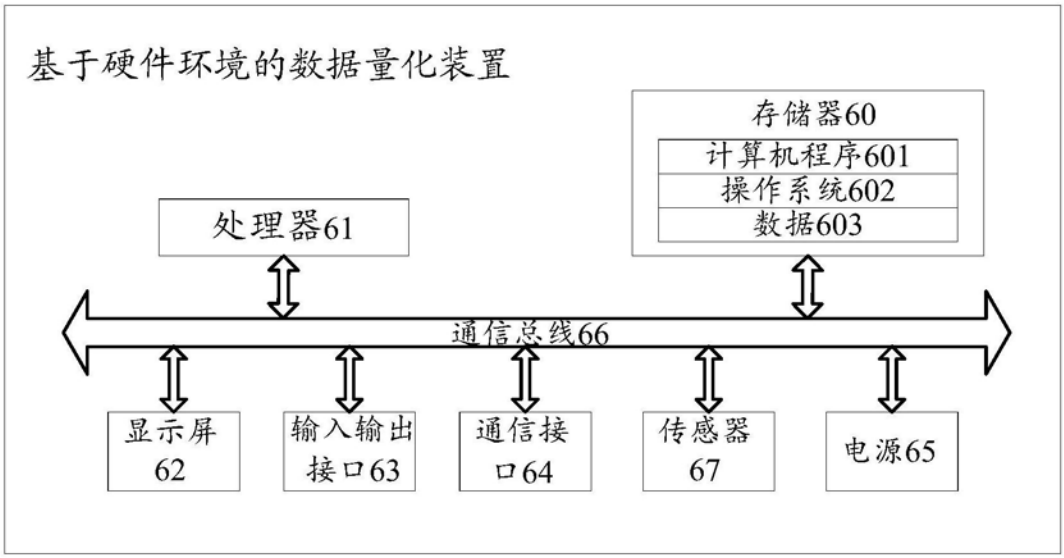

基于硬件环境的数据量化装置
存储器60
计算机程序601
操作系统602
数据603
处理器61
通信总线66
显示屏62
输入输出接口63
通信接口64
传感器67
电源65

图6