(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111325332 B
(45)授权公告日 2023.09.08

(21)申请号 202010098799.8

(22)申请日 2020.02.18

(65)同一申请的已公布的文献号
申请公布号 CN 111325332 A

(43)申请公布日 2020.06.23

(73)专利权人 百度在线网络技术(北京)有限公司
地址 100085 北京市海淀区上地十街10号百度大厦三层

(72)发明人 李强 田超 路阔

(74)专利代理机构 北京清亦华知识产权代理事务所(普通合伙) 11201
专利代理师 王艳斌

(51)Int.Cl.
G06N 3/063(2023.01)
G06N 3/0464(2023.01)
G06F 17/16(2006.01)

(56)对比文件
US 2016342891 A1,2016.11.24
CN 110288086 A,2019.09.27
US 2019325297 A1,2019.10.24
US 2019325296 A1,2019.10.24
US 2017103316 A1,2017.04.13
CN 109190756 A,2019.01.11
US 2018121796 A1,2018.05.03

审查员 李知宇

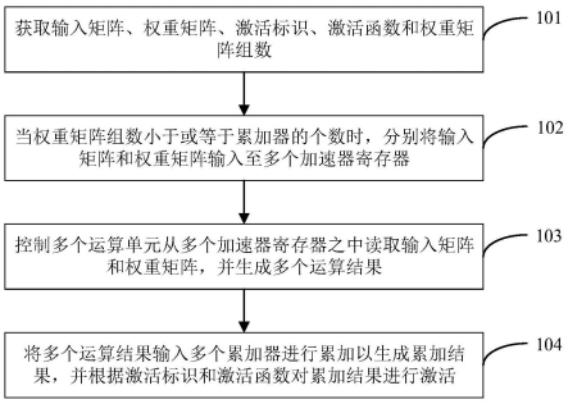
权利要求书2页 说明书10页 附图4页

(54)发明名称

卷积神经网络的处理方法和装置

(57)摘要

本申请公开了一种卷积神经网络的处理方法和装置,涉及计算机技术领域。具体实现方案为:通过获取输入矩阵、权重矩阵、激活标识、激活函数和权重矩阵组数;当权重矩阵组数小于或等于累加器的个数时,分别将输入矩阵和权重矩阵输入至多个加速器寄存器;控制多个运算单元从多个加速器寄存器之中读取输入矩阵和权重矩阵,并生成多个运算结果;以及将多个运算结果输入多个累加器进行累加以生成累加结果,并根据激活标识和激活函数对累加结果进行激活。该方法中对多个累加器生成的累加结果进行激活,与卷积计算的过程均是并行处理的,提高了卷积神经网络的计算效率,进而对因卷积计算效率引起的延迟问题有显著的改善。



CN 111325332 B

1.一种卷积神经网络的处理方法,其特征在于,所述卷积神经网络包括卷积参数寄存器、多个加速器寄存器、与所述多个加速器寄存器分别相连的多个运算单元和与所述多个运算单元分别相连的多个累加器,所述方法包括:

获取输入矩阵、权重矩阵、激活标识、激活函数和权重矩阵组数;

当所述权重矩阵组数小于或等于所述累加器的个数时,分别将所述输入矩阵和所述权重矩阵输入至所述多个加速器寄存器;

控制所述多个运算单元从所述多个加速器寄存器之中读取所述输入矩阵和所述权重矩阵,并生成多个运算结果;以及

将所述多个运算结果输入所述多个累加器进行累加以生成累加结果,并根据所述激活标识和所述激活函数对所述累加结果进行激活;

所述分别将所述输入矩阵和所述权重矩阵输入至所述多个加速器寄存器之前,还包括:

如果所述权重矩阵组数小于或等于所述累加器的个数,则对所述输入矩阵进行转换以生成转换输入矩阵,并对所述权重矩阵进行转换以生成转换权重矩阵;以及

分别将所述转换输入矩阵和所述转换权重矩阵输入所述多个加速器寄存器。

2.如权利要求1所述的卷积神经网络的处理方法,其特征在于,所述对所述输入矩阵进行转换以生成转换输入矩阵,包括:

将所述输入矩阵C*H*W在内存之中转换为H*W*C,其中,C为所述输入矩阵的通道数,H为所述输入矩阵的高,所述W为所述输入矩阵的宽。

3.如权利要求1或2所述的卷积神经网络的处理方法,其特征在于,所述对所述权重矩阵进行转换以生成转换权重矩阵,包括:

将所述权重矩阵数据C*K*K在内存中转换为K*K*C*C',其中,C'为所述权重矩阵组数,K为卷积核尺寸。

4.如权利要求1-2任一项所述的卷积神经网络的处理方法,其特征在于,所述多个加速器寄存器为两个,所述多个运算单元的个数与所述加速器寄存器的大小相对应。

5.如权利要求4所述的卷积神经网络的处理方法,其特征在于,所述运算单元的个数为256个,所述加速器寄存器的大小为256字节,所述累加器的个数为64个。

6.一种卷积神经网络的处理装置,其特征在于,所述卷积神经网络包括卷积参数寄存器、多个加速器寄存器、与所述多个加速器寄存器分别相连的多个运算单元和与所述多个运算单元分别相连的多个累加器,所述装置包括:

获取模块,用于获取输入矩阵、权重矩阵、激活标识、激活函数和权重矩阵组数;

第一输入模块,用于当所述权重矩阵组数小于或等于所述累加器的个数时,分别将所述输入矩阵和所述权重矩阵输入至所述多个加速器寄存器;

生成模块,用于控制所述多个运算单元从所述多个加速器寄存器之中读取所述输入矩阵和所述权重矩阵,并生成多个运算结果;以及

处理模块,用于将所述多个运算结果输入所述多个累加器进行累加以生成累加结果,并根据所述激活标识和所述激活函数对所述累加结果进行激活;

所述装置,还包括:

转换模块,用于如果所述权重矩阵组数小于或等于所述多个累加器的个数,则对所述

输入矩阵进行转换以生成转换输入矩阵，并对所述权重矩阵进行转换以生成转换权重矩阵；以及

第二输入模块，用于分别将所述转换输入矩阵和所述转换权重矩阵输入所述多个加速器寄存器。

7.如权利要求6所述的卷积神经网络的处理装置，其特征在于，所述转换模块，还用于：

将所述输入矩阵C*H*W在内存之中转换为H*W*C，其中，C为所述输入矩阵的通道数，H为所述输入矩阵的高，所述W为所述输入矩阵的宽。

8.如权利要求6或7所述的卷积神经网络的处理装置，其特征在于，所述转换模块，还用于：

将所述权重矩阵数据C*K*K在内存中转换为K*K*C*C'，其中，C'为所述权重矩阵组数，K为卷积核尺寸。

9.如权利要求6-7任一项所述的卷积神经网络的处理装置，其特征在于，所述多个加速器寄存器为两个，所述多个运算单元的个数与所述加速器寄存器的大小相对应。

10.如权利要求9所述的卷积神经网络的处理装置，其特征在于，所述运算单元的个数为256个，所述加速器寄存器的大小为256字节，所述累加器的个数为64个。

11.一种计算机设备，其特征在于，包括：

至少一个处理器；以及

与所述至少一个处理器通信连接的存储器；其中，

所述存储器存储有可被所述至少一个处理器执行的指令，所述指令被所述至少一个处理器执行，以使所述至少一个处理器能够执行权利要求1-5中任一项所述的卷积神经网络的处理方法。

12.一种存储有计算机指令的非瞬时计算机可读存储介质，其特征在于，所述计算机指令用于使所述计算机执行权利要求1-5中任一项所述的卷积神经网络的处理方法。

卷积神经网络的处理方法和装置

技术领域

[0001] 本申请涉及计算机技术领域的卷积神经网络技术领域，尤其涉及一种卷积神经网络的处理方法和装置。

背景技术

[0002] 卷积神经网络是深度学习的关键技术，但是由于卷积神经网络模型具有数量级大、层次复杂、深度大等特点，使用传统处理器单元进行卷积计算时存在效率较低的问题；尤其在语音信号处理等时延要求敏感的场景，卷积计算延迟使得系统实时性要求面临巨大挑战。

[0003] 相关技术中，卷积神经网络在基于模型具体参数处理时，是在所有点完成点积计算后，集中进行激活处理。由于激活处理的速度较慢，导致现有的卷积神经网络存在计算效率低的技术问题。

发明内容

[0004] 本申请第一方面实施例提出了一种卷积神经网络的处理方法，所述卷积神经网络包括卷积参数寄存器、多个加速器寄存器、与所述多个加速器寄存器分别相连的多个运算单元和与所述多个运算单元分别相连的多个累加器，所述方法包括：

[0005] 获取输入矩阵、权重矩阵、激活标识、激活函数和权重矩阵组数；

[0006] 当所述权重矩阵组数小于或等于所述累加器的个数时，分别将所述输入矩阵和所述权重矩阵输入至所述多个加速器寄存器；

[0007] 控制所述多个运算单元从所述多个加速器寄存器之中读取所述输入矩阵和所述权重矩阵，并生成多个运算结果；以及

[0008] 将所述多个运算结果输入所述多个累加器进行累加以生成累加结果，并根据所述激活标识和所述激活函数对所述累加结果进行激活。

[0009] 作为本申请实施例的第一种可能的实现方式，所述分别将所述输入矩阵和所述权重矩阵输入至所述多个加速器寄存器之前，还包括：

[0010] 如果所述权重矩阵组数小于或等于所述多个累加器的个数，则对所述输入矩阵进行转换以生成转换输入矩阵，并对所述权重矩阵进行转换以生成转换权重矩阵；以及

[0011] 分别将所述转换输入矩阵和所述转换权重矩阵输入所述多个加速器寄存器。

[0012] 作为本申请实施例的第二种可能的实现方式，所述对所述输入矩阵进行转换以生成转换输入矩阵，包括：

[0013] 将所述输入矩阵C*H*W在内存之中转换为H*W*C，其中，C为所述输入矩阵的通道数，H为所述输入矩阵的高，所述W为所述输入矩阵的宽。

[0014] 作为本申请实施例的第三种可能的实现方式，所述对所述权重矩阵进行转换以生成转换权重矩阵，包括：

[0015] 将所述权重矩阵数据C*K*K在所述内存中转换为K*K*C*C’，其中，C’为所述权重矩

阵组数,K为卷积核尺寸。

[0016] 作为本申请实施例的第四种可能的实现方式,所述多个加速器寄存器为两个,所述多个运算单元的个数与所述加速器寄存器的大小相对应。

[0017] 作为本申请实施例的第五种可能的实现方式,所述运算单元的个数为256个,所述加速器寄存器的大小为256字节,所述累加器的个数为64个。

[0018] 本申请第二方面实施例提出了一种卷积神经网络的处理装置,所述卷积神经网络包括卷积参数寄存器、多个加速器寄存器、与所述多个加速器寄存器分别相连的多个运算单元和与所述多个运算单元分别相连的多个累加器,所述装置包括:

[0019] 获取模块,用于获取输入矩阵、权重矩阵、激活标识、激活函数和权重矩阵组数;

[0020] 第一输入模块,用于当所述权重矩阵组数小于或等于所述累加器的个数时,分别将所述输入矩阵和所述权重矩阵输入至所述多个加速器寄存器;

[0021] 生成模块,用于控制所述多个运算单元从所述多个加速器寄存器之中读取所述输入矩阵和所述权重矩阵,并生成多个运算结果;以及

[0022] 处理模块,用于将所述多个运算结果输入所述多个累加器进行累加以生成累加结果,并根据所述激活标识和所述激活函数对所述累加结果进行激活。

[0023] 本申请第三方面实施例提出了一种计算机设备,包括:

[0024] 至少一个处理器;以及

[0025] 与所述至少一个处理器通信连接的存储器;其中,

[0026] 所述存储器存储有可被所述至少一个处理器执行的指令,所述指令被所述至少一个处理器执行,以使所述至少一个处理器能够执行第一方面实施例所述的卷积神经网络的处理方法。

[0027] 本申请第四方面实施例提出了一种存储有计算机指令的非瞬时计算机可读存储介质,所述计算机指令用于使所述计算机执行第一方面实施例所述的卷积神经网络的处理方法。

[0028] 上述申请中的一个实施例具有如下优点或有益效果:通过获取输入矩阵、权重矩阵、激活标识、激活函数和权重矩阵组数;当权重矩阵组数小于或等于累加器的个数时,分别将输入矩阵和权重矩阵输入至多个加速器寄存器;控制多个运算单元从多个加速器寄存器之中读取输入矩阵和权重矩阵,并生成多个运算结果;以及将多个运算结果输入多个累加器进行累加以生成累加结果,并根据激活标识和激活函数对累加结果进行激活。该方法中对多个累加器生成的累加结果进行激活,与卷积计算的过程均是并行处理的,相较于相关技术中所有点均完成点积计算后进行激活处理,提高了卷积神经网络的计算效率,进而对因卷积计算效率引起的延迟问题有显著的改善。

[0029] 上述可选方式所具有的其他效果将在下文中结合具体实施例加以说明。

附图说明

[0030] 附图用于更好地理解本方案,不构成对本申请的限定。其中:

[0031] 图1为本申请实施例一提供的卷积神经网络的处理方法的流程示意图;

[0032] 图2为本申请实施例二提供的卷积神经网络的处理方法的流程示意图;

[0033] 图3为本申请实施例提供的一种输入矩阵的结构示例图;

[0034] 图4为本申请实施例提供的一种权重矩阵的结构示例图；

[0035] 图5为本申请实施例提供的一种卷积计算过程的流程示意图；

[0036] 图6为本申请实施例三提供的卷积神经网络的处理装置的结构示意图；

[0037] 图7为本申请实施例三提供的另一种卷积神经网络的处理装置的结构示意图；

[0038] 图8是用来实现本申请实施例的卷积神经网络的处理方法的计算机设备的框图。

具体实施方式

[0039] 以下结合附图对本申请的示范性实施例做出说明，其中包括本申请实施例的各种细节以助于理解，应当将它们认为仅仅是示范性的。因此，本领域普通技术人员应当认识到，可以对这里描述的实施例做出各种改变和修改，而不会背离本申请的范围和精神。同样，为了清楚和简明，以下的描述中省略了对公知功能和结构的描述。

[0040] 下面结合参考附图描述本申请实施例的卷积神经网络的处理方法、装置、计算机设备和存储介质。

[0041] 图1为本申请实施例一提供的卷积神经网络的处理方法的流程示意图。

[0042] 本申请实施例以该卷积神经网络的处理方法被配置于卷积神经网络的处理装置中来举例说明，该卷积神经网络的处理装置可以应用于任一计算机设备中，以使该计算机设备可以执行卷积神经网络的处理功能。

[0043] 其中，计算机设备可以为个人电脑(Personal Computer，简称PC)、云端设备、移动设备等，移动设备例如可以为手机、平板电脑、个人数字助理、穿戴式设备、车载设备等具有各种操作系统的硬件设备。

[0044] 如图1所示，该卷积神经网络的处理方法可以包括以下步骤：

[0045] 步骤101，获取输入矩阵、权重矩阵、激活标识、激活函数和权重矩阵组数。

[0046] 本申请实施例中，卷积神经网络的硬件部分包括卷积参数寄存器、多个加速器寄存器、与多个加速器寄存器分别相连的多个运算单元和与多个运算单元分别相连的多个累加器。

[0047] 其中，卷积参数寄存器，用于存储卷积模型的关键参数，在卷积神经网络进行卷积计算时通过软件进行配置。

[0048] 多个加速器寄存器分别与卷积参数寄存器相连接，例如加速器寄存器可以为2个，每一个加速器寄存器的大小可以为256字节。

[0049] 多个运算单元分别与多个加速器寄存器相连，其中，运算单元的个数可以与加速器寄存器的大小相对应。作为一种示例，加速器寄存器的大小为256字节时，运算单元的个数可以为256个。

[0050] 需要说明的是，运算单元的个数可以与加速器寄存器的大小相对应，加速器寄存器的大小是基于卷积计算性能和卷积神经网络的硬件综合考虑的，以满足最小粒度为int8*int8卷积计算要求。

[0051] 多个累加器分别与多个运算单元相连，用于暂存卷积运算的中间过程结果。作为一种示例，累加器的个数可以为64个，每个累加器的大小为4字节，多个累加器的总容量为256字节，这样必要时累加器也可以当加速器寄存器使用。

[0052] 本申请实施例中，卷积神经网络的软件部分包括卷积参数寄存器设置和模型处理

方式设置两部分。其中，卷积参数包括输入矩阵、权重矩阵、激活标识、激活函数和权重矩阵组数。卷积参数还可以包括矩阵计算类型，例如int8*int8、半精度浮点等。

[0053] 作为一种可能的情况，在进行卷积计算前，通过计算机设备的内嵌汇编指令配置卷积参数，包括输入矩阵、权重矩阵、激活标识、激活函数和权重矩阵组数，从而使得计算机设备获取到输入矩阵、权重矩阵、激活标识、激活函数和权重矩阵组数。

[0054] 例如，输入矩阵的数据排布可以为C*H*W，其中，C为输入矩阵的通道数，H为输入矩阵的高，W为输入矩阵的宽。权重矩阵的数据排布可以为K*K*C*C’，其中，K为卷积核尺寸，C’为权重矩阵组数。

[0055] 步骤102，当权重矩阵组数小于或等于累加器的个数时，分别将输入矩阵和权重矩阵输入至多个加速器寄存器。

[0056] 在一种可能的情况下，获取到的权重矩阵组数小于累加器的个数时，在进行卷积计算时，分别将输入矩阵和权重矩阵输入至多个加速器寄存器。例如，加速器寄存器为两个，将输入矩阵输入第一加速器寄存器，将权重矩阵输入第二加速器寄存器。

[0057] 在另一种可能的情况下，获取到的权重矩阵组数大于或等于累加器的个数时，在进行卷积计算时，采用img2col方式进行矩阵扩展。其中，img2col是在卷积操作中处理矩阵的的算法。img2col算法的基本原理就是把每个卷积核提取成矩阵中的一行元素，再进行矩阵运算。其中，比较常见的卷积核为3*3和5*5。

[0058] 步骤103，控制多个运算单元从多个加速器寄存器之中读取输入矩阵和权重矩阵，并生成多个运算结果。

[0059] 本申请实施例中，在权重矩阵组数小于累加器的个数时，分别将输入矩阵和权重矩阵输入至多个加速器寄存器后，进一步的，控制多个运算单元从多个累加器寄存器的相应位置读取输入矩阵和权重矩阵，进而各个运算单元根据读取到的输入矩阵和权重矩阵进行乘法运算，以生成各个运算单元对应的各运算结果。

[0060] 需要说明的是，当权重矩阵组数小于累加器个数时，可以一次计算出所有的权重矩阵组数个卷积点的部分卷积结果，并将部分卷积结果存于累加器中。

[0061] 步骤104，将多个运算结果输入多个累加器进行累加以生成累加结果，并根据激活标识和激活函数对累加结果进行激活。

[0062] 本申请实施例中，得到多个运算单元对应的运算结果后，将多个运算结果输入多个累加器进行累加计算，以生成累加结果。在完成一批元素的卷积计算后，根据激活标识采用激活函数对累加器中的累加结果进行激活处理，最后将激活结果从累加器写回对应的地址。

[0063] 需要说明的是，根据激活标识对多个累加器中的累加结果进行激活处理的过程是并行处理过程，并且某点的激活处理和其他点的卷积计算也是并行进行的，从而提高了卷积神经网络的处理效率，降低了卷积计算延迟。

[0064] 本申请实施例中的卷积神经网络的处理方法，通过获取输入矩阵、权重矩阵、激活标识、激活函数和权重矩阵组数；当权重矩阵组数小于或等于累加器的个数时，分别将输入矩阵和权重矩阵输入至多个加速器寄存器；控制多个运算单元从多个加速器寄存器之中读取输入矩阵和权重矩阵，并生成多个运算结果；以及将多个运算结果输入多个累加器进行累加以生成累加结果，并根据激活标识和激活函数对累加结果进行激活。该方法中对多个

累加器生成的累加结果进行激活，与卷积计算的过程均是并行处理的，相较于相关技术中所有点均完成点积计算后进行激活处理，提高了卷积神经网络的计算效率，进而对因卷积计算效率引起的延迟问题有显著的改善。

[0065] 在上述实施例的基础上，在上述步骤102中，分别将输入矩阵和权重矩阵输入至多个加速器寄存器之前，确定权重矩阵组数小于或等于多个累加器的个数后，还可以对输入矩阵进行转换以生成转换输入矩阵，并对权重矩阵进行转换以生成转换权重矩阵，进而，分别将转换输入矩阵和转换权重矩阵输入多个加速器寄存器。下面结合图2对上述过程进行详细介绍，图2为本申请实施例二提供的卷积神经网络的处理方法的流程示意图。

[0066] 如图2所示，该卷积神经网络的处理方法，还可以包括以下步骤：

[0067] 步骤201，获取权重矩阵组数。

[0068] 本申请实施例中，权重矩阵组数可以为用户预先通过内嵌汇编指令配置的，从而使得计算机设备获取到权重矩阵组数。

[0069] 步骤202，如果权重矩阵组数小于或等于累加器的个数，则对输入矩阵进行转换以生成转换输入矩阵，并对权重矩阵进行转换以生成转换权重矩阵。

[0070] 本申请实施例中，当获取到的权重矩阵组数小于或等于累加器的个数时，则对输入矩阵进行转换，以生成转换输入矩阵。例如，累加器的个数可以为64个。

[0071] 作为一种可能的实现方式，假设输入矩阵的数据排布方式为C*H*W，可以将输入矩阵C*H*W在内存之中转换为转换输入矩阵H*W*C，其中，C为输入矩阵的通道数，H为输入矩阵的高，W为输入矩阵的宽。

[0072] 作为一种实例，输入矩阵可以如图3中所示，图3中箭头表示存储的输入矩阵在通道维度上是连续的。

[0073] 需要说明的是，将输入矩阵在内存之中转换为转换输入矩阵，是发送矩阵数据变换指令，由卷积神经网络的硬件完成的。

[0074] 由此，通过将输入矩阵在内存中的排布方式变换为H*W*C，这样内存中输入矩阵的通道在内存，便于计算出一个点或者多个点的点积，在其中一个点积计算完成后，便可进行激活。其中，激活的过程可以和其他点的点积计算并行处理，从而提高了卷积神经网络的处理速率。

[0075] 本申请实施例中，权重矩阵数据的排列方式可以为C*K*K，在有C’组权重矩阵数据时，需要将权重矩阵数据在内存中的数据排布转换为K*K*C*C’，其中，C’为权重矩阵组数，K为卷积核尺寸。

[0076] 作为一种示例，如图4所示，为C’组权重矩阵C*K*K，其中，C’为权重矩阵组数。权重矩阵在内存中的数据排布由图4中的C*K*K，转为为徒5中的K*K*C*C’。在进行卷积计算时，在输入特征图H*W*C1时，对应的权重矩阵为K*K*(C1*C’)，在输入特征图H*W*C2时，对应的权重矩阵为K*K*(C2*C’)。

[0077] 步骤203，分别将转换输入矩阵和转换权重矩阵输入多个加速器寄存器。

[0078] 本申请实施例中，完成对输入矩阵进行转换以生成转换输入矩阵，并对权重矩阵进行转换以生成转换权重矩阵后，分别将转换输入矩阵和转换权重矩阵输入多个加速器寄存器。进而，控制多个运算单元从多个加速器寄存器之中读取输入矩阵和权重矩阵，并生成多个运算结果，以将多个运算结果输入多个累加器进行累加以生成累加结果，并根据激活

标识和激活函数对累加结果进行激活。

[0079] 作为一种可能的情况，如图5所示，若转换权重矩阵K*K*C*C'中，C x*C'小于或等于字节阈值时，其中，字节阈值为256字节，则从输入矩阵地址偏移0*C字节、加载C字节个数据至第一加速器寄存器中，并从转换权重矩阵地址偏移0*(CxC')字节、加载Cx*C'个字节数据至第二加速器寄存器，进而，将多个运算单元进行计算生成的多个运算结果存储至累加器1至C'；进一步的，输入矩阵地址偏移1*C字节、加载C字节个数据至第一加速器寄存器中，转换权重矩阵地址偏移1*(CxC')字节、加载CxC'个字节数据至第二加速器寄存器，进而，将多个运算单元进行计算生成的多个运算结果存储至累加器1至C'；依照上述步骤进行k次，在第k次计算时，输入矩阵地址偏移(k-1)*C字节、加载C字节个数据至第一加速器寄存器中，转换权重矩阵地址偏移(k-1)*(CxC')字节、加载CxC'个字节数据至第二加速器寄存器，将多个运算单元进行计算生成的多个运算结果存储至累加器1至C'；在进行n*k+m次运算时，输入矩阵地址偏移n*C+(m-1)*C字节、加载C字节个数据至第一加速器寄存器中，转换权重矩阵地址偏移(n*k+m-1)*(CxC')字节、加载CxC'个字节数据至第二加速器寄存器中，将多个运算单元进行计算生成的多个运算结果存储至累加器1至C'；依次，在完成k*k次运算后，C'个元素的卷积计算完成，依据要求进行激活处理后将结果从累加器写回对应地址。

[0080] 沿着图5所示卷积计算K*K*C*C'转换权重矩阵移动顺序，进行下一组C'个元素的卷积计算；计算下一组时输入矩阵起始地址偏移C字节，重复上述步骤。依次重复上述过程，直至完成H*W*C'元素卷积处理。

[0081] 作为另一种可能的情况，若转换权重矩阵K*K*C*C'中，C x*C'大于字节阈值时，其中，字节阈值为256字节，需要将C x*C'拆分成多次完成卷积计算过程。对照于CxC'小于等于256字节时的第一次计算为例，CxC'计算需要拆分成多次完成，如设(256/C')*C'＝len，从输入矩阵地址偏移0*C字节、加载256/C'字节个数据至第一加速器寄存器中，然后从转换权重矩阵地址偏移0*(CxC')字节、加载len个数据至第二加速器寄存器中，然后将多个运算单元进行计算生成的多个运算结果存储至累加器1至C'。进一步的，从输入矩阵地址偏移0*C+256/C'字节、加载256/C'字节个数据至第一加速器寄存器中，然后从转换权重矩阵地址偏移0*(CxC')+len字节、加载len个数据至第二加速器寄存器中，然后将多个运算单元进行计算生成的多个运算结果存储至累加器1至C'；直至完成CxC'数据量的处理。

[0082] 本申请实施例中，字节阈值设置为256字节时，是基于加速器寄存器的大小考虑的。

[0083] 需要说明的是，为了最大长度挖掘卷积神经网络计算的并行性，输入矩阵转换以生成转换输入矩阵是由卷积神经网络的硬件完成的，软件中需要进行相应操作时，发送一条内嵌自定义的汇编指令即可。往加速器寄存器加载数据、运算单元进行计算以及从多个累加器导出数据，均需要自定内嵌汇编指令完成。

[0084] 本申请实施例的卷积神经网络的处理方法，通过获取权重矩阵组数，确定权重矩阵组数小于或等于累加器的个数，则对输入矩阵进行转换以生成转换输入矩阵，并对权重矩阵进行转换以生成转换权重矩阵，进而，分别将转换输入矩阵和转换权重矩阵输入多个加速器寄存器。由此，在卷积神经网络进行卷积计算过程中，将卷积结果加载至加速器寄存器进行激活操作，极大了提高了卷积计算效率，降低卷积计算延迟。

[0085] 为了实现上述实施例，本申请实施例提出了一种卷积神经网络的处理装置。

[0086] 图6为本申请实施例三提供的卷积神经网络的处理装置的结构示意图。

[0087] 其中，卷积神经网络包括卷积参数寄存器、多个加速器寄存器、与多个加速器寄存器分别相连的多个运算单元和与多个运算单元分别相连的多个累加器。

[0088] 如图6所示，该卷积神经网络的处理装置300，可以包括：获取模块310、第一输入模块320、生成模块330以及处理模块340。

[0089] 其中，获取模块310，用于获取输入矩阵、权重矩阵、激活标识、激活函数和权重矩阵组数。

[0090] 第一输入模块320，用于当权重矩阵组数小于或等于累加器的个数时，分别将输入矩阵和权重矩阵输入至多个加速器寄存器。

[0091] 生成模块330，用于控制多个运算单元从多个加速器寄存器之中读取输入矩阵和权重矩阵，并生成多个运算结果。

[0092] 处理模块340，用于将多个运算结果输入多个累加器进行累加以生成累加结果，并根据激活标识和激活函数对累加结果进行激活。

[0093] 作为一种可能的情况，参见图7，该卷积神经网络的处理装置300，还可以包括：

[0094] 转换模块350，用于如果权重矩阵组数小于或等于累加器的个数，则对输入矩阵进行转换以生成转换输入矩阵，并对权重矩阵进行转换以生成转换权重矩阵。

[0095] 第二输入模块360，用于分别将转换输入矩阵和转换权重矩阵输入多个加速器寄存器。

[0096] 作为另一种可能的情况，转换模块360，还可以用于：

[0097] 将输入矩阵C*H*W在内存之中转换为H*W*C，其中，C为输入矩阵的通道数，H为输入矩阵的高，W为输入矩阵的宽。

[0098] 作为另一种可能的情况，转换模块360，还可以用于：

[0099] 将权重矩阵数据C*K*K在内存中转换为K*K*C*C’，其中，C’为权重矩阵组数，K为卷积核尺寸。

[0100] 作为另一种可能的情况，多个加速器寄存器为两个，多个运算单元的个数与加速器寄存器的大小相对应。

[0101] 作为另一种可能的情况，运算单元的个数为256个，加速器寄存器的大小为256字节，累加器的个数为64个。

[0102] 本申请实施例中的卷积神经网络的处理装置，通过获取输入矩阵、权重矩阵、激活标识、激活函数和权重矩阵组数；当权重矩阵组数小于或等于累加器的个数时，分别将输入矩阵和权重矩阵输入至多个加速器寄存器；控制多个运算单元从多个加速器寄存器之中读取输入矩阵和权重矩阵，并生成多个运算结果；以及将多个运算结果输入多个累加器进行累加以生成累加结果，并根据激活标识和激活函数对累加结果进行激活。该方法中对多个累加器生成的累加结果进行激活，与卷积计算的过程均是并行处理的，相较于相关技术中所有点均完成点积计算后进行激活处理，提高了卷积神经网络的计算效率，进而对因卷积计算效率引起的延迟问题有显著的改善。

[0103] 根据本申请的实施例，本申请还提供了一种计算机设备和一种可读存储介质。

[0104] 如图8所示，是根据本申请实施例的卷积神经网络的处理方法的计算机设备的框图。计算机设备旨在表示各种形式的数字计算机，诸如，膝上型计算机、台式计算机、工作

台、个人数字助理、服务器、刀片式服务器、大型计算机、和其它适合的计算机。计算机设备还可以表示各种形式的移动装置，诸如，个人数字处理、蜂窝电话、智能电话、可穿戴设备和其它类似的计算装置。本文所示的部件、它们的连接和关系、以及它们的功能仅仅作为示例，并且不意在限制本文中描述的和/或者要求的本申请的实现。

[0105] 如图8所示，该计算机设备包括：一个或多个处理器501、存储器502，以及用于连接各部件的接口，包括高速接口和低速接口。各个部件利用不同的总线互相连接，并且可以被安装在公共主板上或者根据需要以其它方式安装。处理器可以对在计算机设备内执行的指令进行处理，包括存储在存储器中或者存储器上以在外部输入/输出装置(诸如，耦合至接口的显示设备)上显示GUI的图形信息的指令。在其它实施方式中，若需要，可以将多个处理器和/或多条总线与多个存储器和多个存储器一起使用。同样，可以连接多个计算机设备，各个设备提供部分必要的操作(例如，作为服务器阵列、一组刀片式服务器、或者多处理器系统)。图8中以一个处理器501为例。

[0106] 存储器502即为本申请所提供的非瞬时计算机可读存储介质。其中，所述存储器存储有可由至少一个处理器执行的指令，以使所述至少一个处理器执行本申请所提供的卷积神经网络的处理方法。本申请的非瞬时计算机可读存储介质存储计算机指令，该计算机指令用于使计算机执行本申请所提供的卷积神经网络的处理方法。

[0107] 存储器502作为一种非瞬时计算机可读存储介质，可用于存储非瞬时软件程序、非瞬时计算机可执行程序以及模块，如本申请实施例中的卷积神经网络的处理方法对应的程序指令/模块(例如，附图6所示的获取模块310、第一输入模块320、生成模块330以及处理模块340)。处理器501通过运行存储在存储器502中的非瞬时软件程序、指令以及模块，从而执行服务器的各种功能应用以及数据处理，即实现上述方法实施例中的卷积神经网络的处理方法。

[0108] 存储器502可以包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需要的应用程序；存储数据区可存储根据卷积神经网络的处理的计算机设备的使用所创建的数据等。此外，存储器502可以包括高速随机存取存储器，还可以包括非瞬时存储器，例如至少一个磁盘存储器件、闪存器件、或其他非瞬时固态存储器件。在一些实施例中，存储器502可选包括相对于处理器501远程设置的存储器，这些远程存储器可以通过网络连接至卷积神经网络的处理的计算机设备。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

[0109] 卷积神经网络的处理方法的计算机设备还可以包括：输入装置503和输出装置504。处理器501、存储器502、输入装置503和输出装置504可以通过总线或者其他方式连接，图8中以通过总线连接为例。

[0110] 输入装置503可接收输入的数字或字符信息，以及产生与卷积神经网络的处理的计算机设备的用户设置以及功能控制有关的键信号输入，例如触摸屏、小键盘、鼠标、轨迹板、触摸板、指示杆、一个或者多个鼠标按钮、轨迹球、操纵杆等输入装置。输出装置504可以包括显示设备、辅助照明装置(例如，LED)和触觉反馈装置(例如，振动电机)等。该显示设备可以包括但不限于，液晶显示器(LCD)、发光二极管(LED)显示器和等离子体显示器。在一些实施方式中，显示设备可以是触摸屏。

[0111] 此处描述的系统和技术的各种实施方式可以在数字电子电路系统、集成电路系

统、专用ASIC(专用集成电路)、计算机硬件、固件、软件、和/或它们的组合中实现。这些各种实施方式可以包括:实施在一个或者多个计算机程序中,该一个或者多个计算机程序可在包括至少一个可编程处理器的可编程系统上执行和/或解释,该可编程处理器可以是专用或者通用可编程处理器,可以从存储系统、至少一个输入装置、和至少一个输出装置接收数据和指令,并且将数据和指令传输至该存储系统、该至少一个输入装置、和该至少一个输出装置。

[0112] 这些计算程序(也称作程序、软件、软件应用、或者代码)包括可编程处理器的机器指令,并且可以利用高级过程和/或面向对象的编程语言、和/或汇编/机器语言来实施这些计算程序。如本文使用的,术语"机器可读介质"和"计算机可读介质"指的是用于将机器指令和/或数据提供给可编程处理器的任何计算机程序产品、设备、和/或装置(例如,磁盘、光盘、存储器、可编程逻辑装置(PLD)),包括,接收作为机器可读信号的机器指令的机器可读介质。术语"机器可读信号"指的是用于将机器指令和/或数据提供给可编程处理器的任何信号。

[0113] 为了提供与用户的交互,可以在计算机上实施此处描述的系统和技术,该计算机具有:用于向用户显示信息的显示装置(例如,CRT(阴极射线管)或者LCD(液晶显示器)监视器);以及键盘和指向装置(例如,鼠标或者轨迹球),用户可以通过该键盘和该指向装置来将输入提供给计算机。其它种类的装置还可以用于提供与用户的交互;例如,提供给用户的反馈可以是任何形式的传感反馈(例如,视觉反馈、听觉反馈、或者触觉反馈);并且可以用任何形式(包括声输入、语音输入或者、触觉输入)来接收来自用户的输入。

[0114] 可以将此处描述的系统和技术实施在包括后台部件的计算系统(例如,作为数据服务器)、或者包括中间件部件的计算系统(例如,应用服务器)、或者包括前端部件的计算系统(例如,具有图形用户界面或者网络浏览器的用户计算机,用户可以通过该图形用户界面或者该网络浏览器来与此处描述的系统和技术的实施方式交互)、或者包括这种后台部件、中间件部件、或者前端部件的任何组合的计算系统中。可以通过任何形式或者介质的数字数据通信(例如,通信网络)来将系统的部件相互连接。通信网络的示例包括:局域网(LAN)、广域网(WAN)和互联网。

[0115] 计算机系统可以包括客户端和服务器。客户端和服务器一般远离彼此并且通常通过通信网络进行交互。通过在相应的计算机上运行并且彼此具有客户端-服务器关系的计算机程序来产生客户端和服务器的关系。

[0116] 根据本申请实施例的技术方案,通过获取输入矩阵、权重矩阵、激活标识、激活函数和权重矩阵组数;当权重矩阵组数小于或等于累加器的个数时,分别将输入矩阵和权重矩阵输入至多个加速器寄存器;控制多个运算单元从多个加速器寄存器之中读取输入矩阵和权重矩阵,并生成多个运算结果;以及将多个运算结果输入多个累加器进行累加以生成累加结果,并根据激活标识和激活函数对累加结果进行激活。该方法中对多个累加器生成的累加结果进行激活,与卷积计算的过程均是并行处理的,相较于相关技术中所有点均完成点积计算后进行激活处理,提高了卷积神经网络的计算效率,进而对因卷积计算效率引起的延迟问题有显著的改善。

[0117] 应该理解,可以使用上面所示的各种形式的流程,重新排序、增加或删除步骤。例如,本发申请中记载的各步骤可以并行地执行也可以顺序地执行也可以不同的次序执行,

只要能够实现本申请公开的技术方案所期望的结果，本文在此不进行限制。

[0118] 上述具体实施方式，并不构成对本申请保护范围的限制。本领域技术人员应该明白的是，根据设计要求和其他因素，可以进行各种修改、组合、子组合和替代。任何在本申请的精神和原则之内所作的修改、等同替换和改进等，均应包含在本申请保护范围之内。

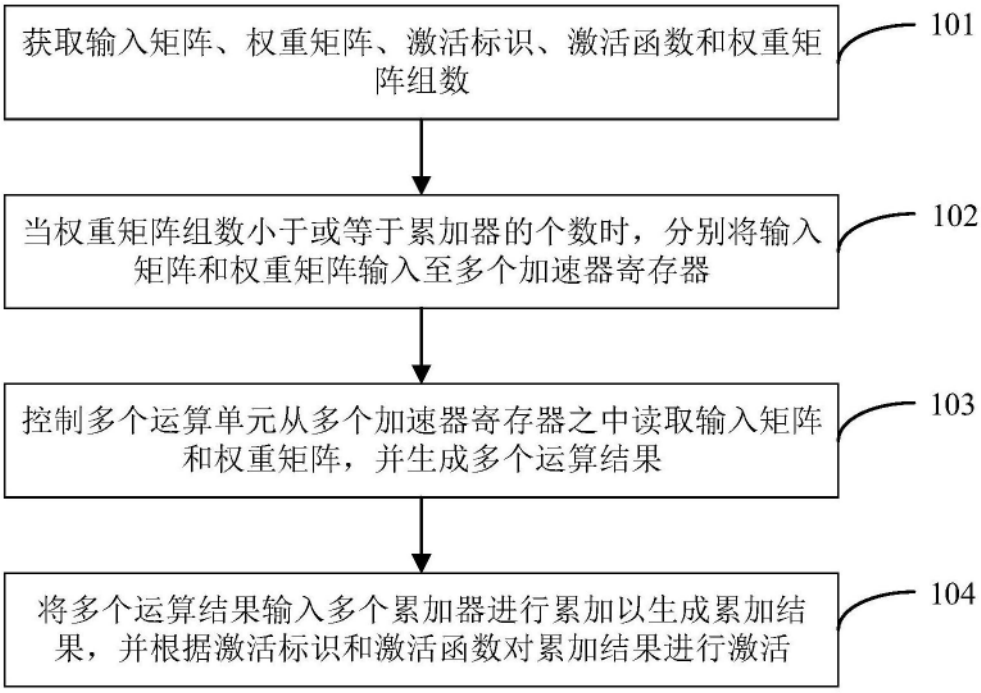

获取输入矩阵、权重矩阵、激活标识、激活函数和权重矩阵组数
101
当权重矩阵组数小于或等于累加器的个数时，分别将输入矩阵和权重矩阵输入至多个加速器寄存器
102
控制多个运算单元从多个加速器寄存器之中读取输入矩阵和权重矩阵，并生成多个运算结果
103
将多个运算结果输入多个累加器进行累加以生成累加结果，并根据激活标识和激活函数对累加结果进行激活
104

图1

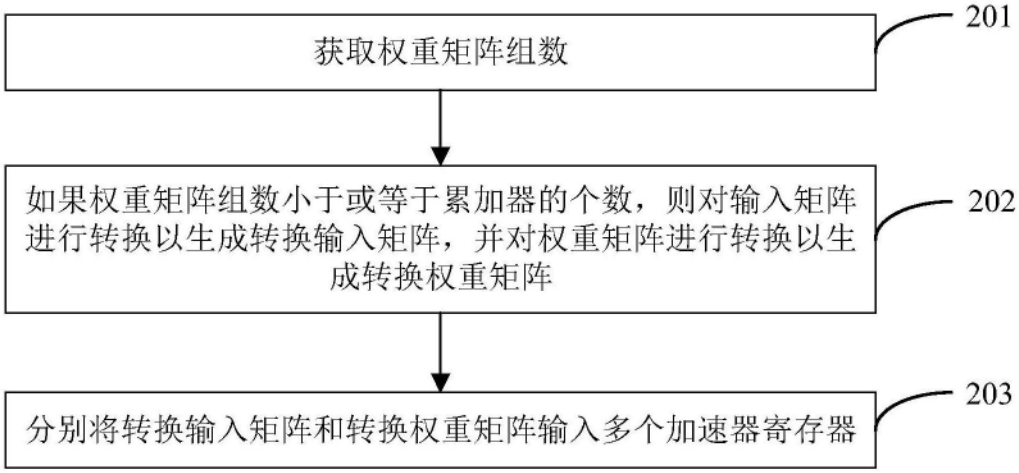

获取权重矩阵组数
201
如果权重矩阵组数小于或等于累加器的个数，则对输入矩阵进行转换以生成转换输入矩阵，并对权重矩阵进行转换以生成转换权重矩阵
202
分别将转换输入矩阵和转换权重矩阵输入多个加速器寄存器
203

图2

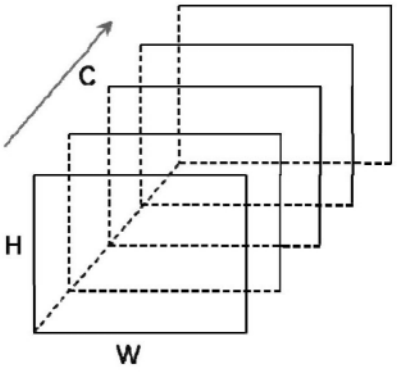

C
H
W

图3

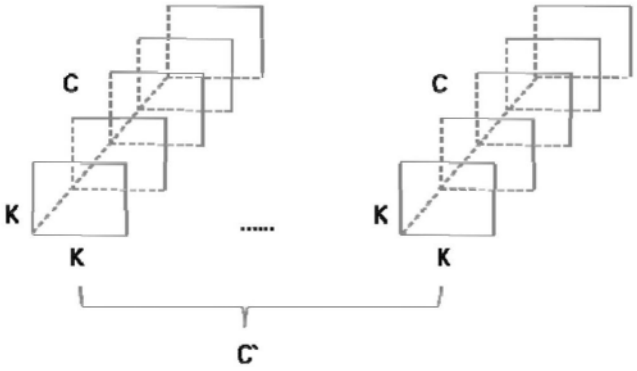

C
K
K
.....
C
K
K
C`

图4

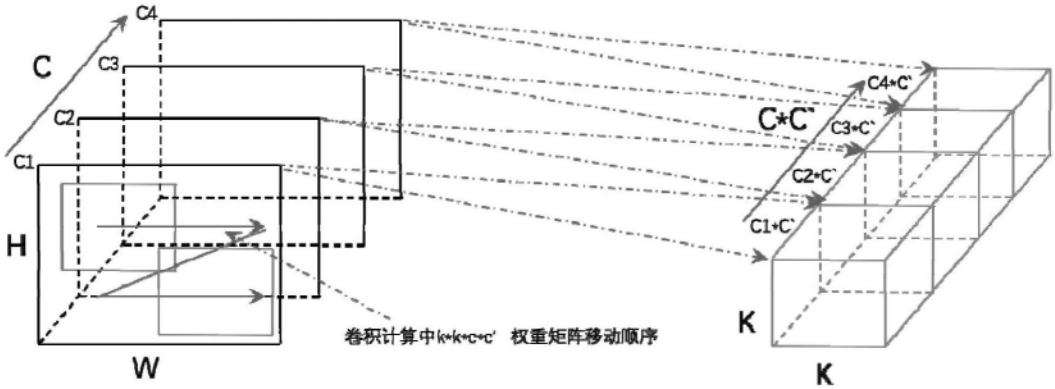

C4
C
C3
C2
C1
H
W
卷积计算中k*k*c*c' 权重矩阵移动顺序
C4*C`
C*C`
C3*C`
C2*C`
C1*C`
K
K

图5

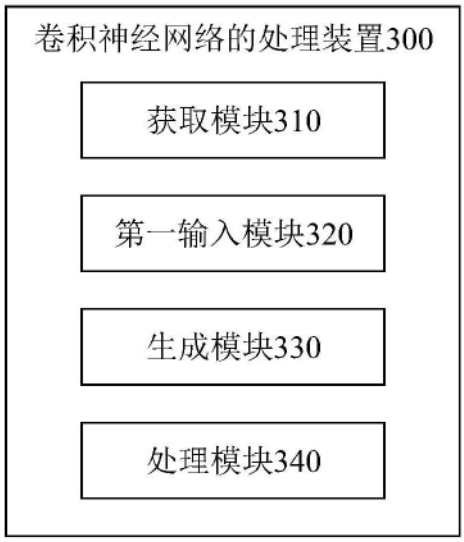

卷积神经网络的处理装置300
获取模块310
第一输入模块320
生成模块330
处理模块340

图6

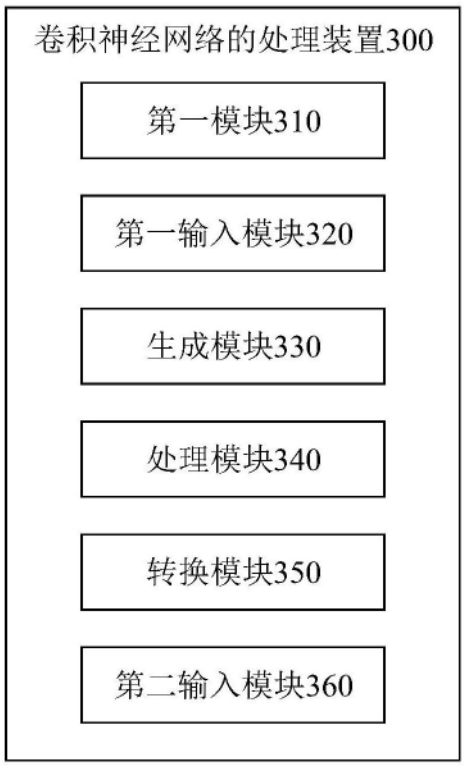

卷积神经网络的处理装置300
第一模块310
第一输入模块320
生成模块330
处理模块340
转换模块350
第二输入模块360

图7

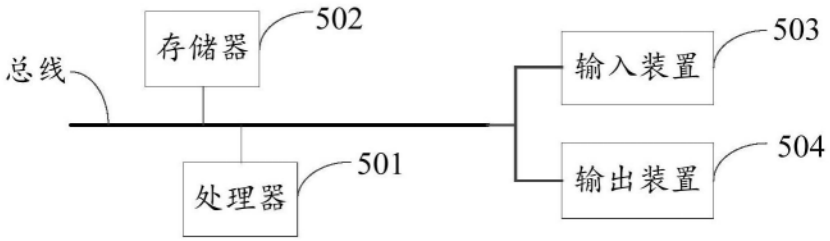


图8