(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 108491359 B
(45)授权公告日 2019.12.24

(21)申请号 201810236769.1

(22)申请日 2016.04.22

(65)同一申请的已公布的文献号
申请公布号 CN 108491359 A

(43)申请公布日 2018.09.04

(62)分案原申请数据
201610258546.6 2016.04.22

(73)专利权人 北京中科寒武纪科技有限公司
地址 100191 北京市海淀区科学院南路6号科研综合楼644室

(72)发明人 刘少礼 张潇 陈云霁 陈天石

(74)专利代理机构 北京华进京联知识产权代理有限公司 11606
代理人 王程

(51)Int.Cl.
G06F 17/16(2006.01)

审查员 邓欣

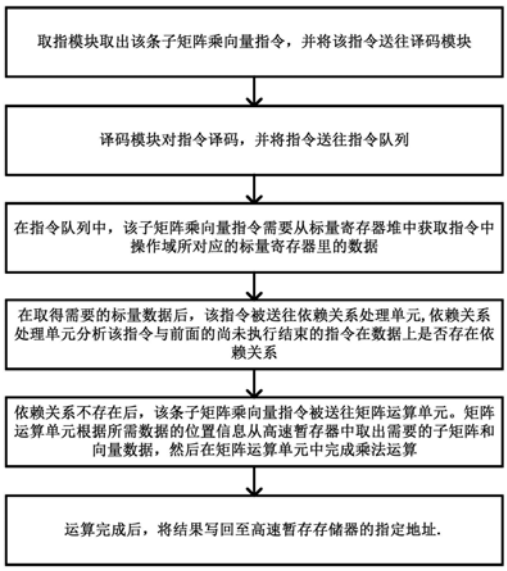
权利要求书3页 说明书12页 附图4页

(54)发明名称

子矩阵运算装置及方法

(57)摘要

本发明提供了一种子矩阵运算装置及方法，上述方法包括如下步骤：获取子矩阵运算指令，子矩阵运算指令包括张量运算指令、子矩阵乘向量指令、向量乘子矩阵指令、子矩阵加法指令、子矩阵减法指令以及子矩阵乘法指令中的至少一种；根据子矩阵运算指令分别从寄存器单元中获取第一子矩阵信息和第二子矩阵信息；根据第一子矩阵信息从存储单元中获取第一子矩阵数据，根据第二子矩阵信息从存储单元中获取第二子矩阵数据；根据第一子矩阵数据和第二子矩阵数据进行子矩阵运算，获得子矩阵运算结果。本发明的子矩阵运算装置及方法，使得子矩阵运算过程中可以更加灵活有效地支持不同宽度的数据，提高了张量运算及子矩阵加减乘除运算等运算的运算效率。



CN 108491359 B

1.一种子矩阵运算方法,其特征在于,所述方法包括如下步骤:

获取子矩阵运算指令,其中,所述子矩阵运算指令包括张量运算指令、子矩阵乘向量指令、向量乘子矩阵指令、子矩阵加法指令、子矩阵减法指令以及子矩阵乘法指令中的至少一种;

根据所述子矩阵运算指令分别从寄存器单元中获取第一子矩阵信息和第二子矩阵信息,所述第一子矩阵信息包括所述第一子矩阵数据在存储单元中的起始地址、所述第一子矩阵数据的行宽、所述第一子矩阵数据的列宽以及所述第一子矩阵数据的行间隔,其中,所述第一子矩阵数据的行间隔是指所述第一子矩阵数据相邻两行间,上一行的行末数据到下一行的行首数据的数据间隔;

根据所述第一子矩阵信息从存储单元中获取第一子矩阵数据,根据所述第二子矩阵信息从所述存储单元中获取第二子矩阵数据,第一子矩阵为二维矩阵的二维子矩阵,所述根据所述第一子矩阵信息从存储单元中获取第一子矩阵数据,包括:从所述第一子矩阵数据在存储单元中的起始地址开始,每读取所述第一子矩阵数据的行宽个数据后跳过所述第一子矩阵数据的行间隔个数据再读取所述第一子矩阵数据的行宽个数据,重复所述第一子矩阵数据的列宽次,得到所述第一子矩阵数据;

根据所述第一子矩阵数据和所述第二子矩阵数据进行子矩阵运算,获得子矩阵运算结果。

2.根据权利要求1所述的方法,其特征在于,所述第二子矩阵信息包括向量地址及向量长度;

根据所述第一子矩阵数据和所述第二子矩阵数据进行子矩阵运算,获得子矩阵运算结果的步骤包括:

将所述第一子矩阵数据作为被乘数,将所述第二子矩阵数据作为乘数进行子矩阵乘向量运算,获得子矩阵乘向量运算结果;

或者,将所述第一子矩阵数据作为乘数,将所述第二子矩阵数据作为被乘数进行向量乘子矩阵运算,获得向量乘子矩阵运算结果。

3.根据权利要求1所述的方法,其特征在于,所述第二子矩阵信息包括所述第二子矩阵数据在所述存储单元中的起始地址、所述第二子矩阵数据的行宽、所述第二子矩阵数据的列宽以及所述第二子矩阵数据的行间隔,其中,所述第二子矩阵数据的行间隔是指所述第二子矩阵数据相邻两行间,上一行的行末数据到下一行的行首数据的数据间隔;

根据所述第一子矩阵数据和所述第二子矩阵数据进行子矩阵运算,获得子矩阵运算结果的步骤包括:

根据所述第一子矩阵数据和所述第二子矩阵数据进行矩阵加法运算或减法运算。

4.根据权利要求1所述的方法,其特征在于,所述第二子矩阵信息包括所述第二子矩阵数据在所述存储单元中的起始地址、所述第二子矩阵数据的行宽、所述第二子矩阵数据的列宽及所述第二子矩阵数据的行间隔,其中,所述第二子矩阵数据的行间隔是指所述第二子矩阵数据相邻两行间,上一行的行末数据到下一行的行首数据的数据间隔;

根据所述第一子矩阵数据和所述第二子矩阵数据进行子矩阵运算,获得子矩阵运算结果的步骤包括:

根据所述第一子矩阵数据和所述第二子矩阵数据进行对位乘法运算,获得子矩阵乘法

运算结果。

5.根据权利要求1所述的方法，其特征在于，所述第二子矩阵信息包括所述第二子矩阵数据在所述存储单元中的起始地址、所述第二子矩阵数据的行宽、所述第二子矩阵数据的列宽及所述第二子矩阵数据的行间隔，其中，所述第二子矩阵数据的行间隔是指所述第二子矩阵数据相邻两行之间，上一行的行末数据到下一行的行首数据的数据间隔；

根据所述第一子矩阵数据和所述第二子矩阵数据进行子矩阵运算，获得子矩阵运算结果的步骤包括：

根据所述第一子矩阵数据和所述第二子矩阵数据进行张量运算，获得张量运算结果。

6.根据权利要求1-5任一项所述的子矩阵运算方法，其特征在于，所述子矩阵运算指令包括操作码和至少一个操作域，其中，所述操作码用于指示所述子矩阵运算指令的功能，操作域用于指示所述子矩阵运算指令的数据信息；

所述子矩阵运算指令的数据信息包括所述寄存器单元的编号，从而能够根据寄存器单元的编号访问对应的寄存器单元，获取所述第一子矩阵信息和所述第二子矩阵信息。

7.根据权利要求1-5任一项所述的子矩阵运算方法，其特征在于，所述方法还包括如下步骤：

对获取的子矩阵运算指令进行译码；

判断所述子矩阵运算指令与前一子矩阵运算指令是否访问相同的子矩阵数据，若是，则将所述子矩阵运算指令存储在指令队列中，等待前一子矩阵运算指令执行完毕后，再执行根据所述子矩阵运算指令分别从寄存器单元中获取第一子矩阵信息和所述第二子矩阵信息步骤。

8.一种子矩阵运算装置，其特征在于，用于根据子矩阵运算指令从矩阵数据中获取子矩阵数据，并根据所述子矩阵数据执行子矩阵运算，所述装置包括：

存储单元，用于存储矩阵数据；

寄存器单元，用于存储子矩阵信息；

子矩阵运算单元，用于获取子矩阵运算指令，根据所述子矩阵运算指令分别从所述寄存器单元中获取第一子矩阵信息和第二子矩阵信息；根据所述第一子矩阵信息从所述存储单元中获取第一子矩阵数据，根据所述第二子矩阵信息从所述存储单元中获取第二子矩阵数据；并根据所述第一子矩阵数据和所述第二子矩阵数据进行子矩阵运算，获得子矩阵运算结果，所述第一子矩阵信息包括第一子矩阵数据在所述存储单元中的起始地址、第一子矩阵数据的行宽、第一子矩阵数据的列宽及第一子矩阵数据的行间隔，其中，所述第一子矩阵数据的行间隔是指第一子矩阵数据相邻两行间，上一行的行末数据到下一行的行首数据的数据间隔，第一子矩阵为二维矩阵的二维子矩阵，所述子矩阵运算单元，具体用于：从所述第一子矩阵数据在存储单元中的起始地址开始，每读取所述第一子矩阵数据的行宽个数据后跳过所述第一子矩阵数据的行间隔个数据再读取所述第一子矩阵数据的行宽个数据，重复所述第一子矩阵数据的列宽次，得到所述第一子矩阵数据；

其中，所述子矩阵运算指令包括张量运算指令、子矩阵乘向量指令、向量乘子矩阵指令、子矩阵加法指令、子矩阵减法指令以及子矩阵乘法指令。

9.根据权利要求8所述的子矩阵运算装置，其特征在于，所述子矩阵运算指令为子矩阵乘向量指令或向量乘子矩阵运算指令；所述第二子矩阵信息包括向量地址及向量长度。

10.根据权利要求8所述的子矩阵运算装置，其特征在于，所述子矩阵运算指令为张量运算指令、子矩阵加法指令、子矩阵减法指令或子矩阵对位乘法指令；

所述第二子矩阵信息包括所述第二子矩阵数据在所述存储单元中的起始地址、所述第二子矩阵数据的行宽、所述第二子矩阵数据的列宽、所述第二子矩阵数据的行间隔，其中，所述第二子矩阵数据的行间隔是指所述第二子矩阵数据相邻两行间，上一行的行末数据到下一行的行首数据的数据间隔。

11.根据权利要求8-10任一项所述的子矩阵运算装置，其特征在于，所述装置还包括用于获取所述子矩阵运算指令，并将所述子矩阵运算指令进行处理的指令处理单元；所述指令处理单元包括：

取址模块，用于获取所述子矩阵运算指令；

译码模块，用于对获取的所述子矩阵运算指令进行译码；

指令队列，用于对译码后的所述子矩阵运算指令进行顺序存储；

依赖关系处理单元，用于在所述子矩阵运算单元获取所述子矩阵运算指令之前，判断所述子矩阵运算指令与前一子矩阵运算指令是否访问相同的子矩阵数据，若是，则将所述子矩阵运算指令存储在指令队列中，等待前一子矩阵运算指令执行完毕后，再根据所述子矩阵运算指令分别获取第一子矩阵信息和第二子矩阵信息。

12.根据权利要求8-10任一项所述的子矩阵运算装置，其特征在于，所述存储单元还用于存储子矩阵运算结果；

所述装置还包括输入输出单元，所述输入输出单元用于将矩阵数据存储至所述存储单元，所述输入输出单元还用于从所述存储单元中获取所述子矩阵运算结果。

13.根据权利要求8-10任一项所述的子矩阵运算装置，其特征在于，所述存储单元为高速暂存存储器。

14.根据权利要求8-10任一项所述的子矩阵运算装置，其特征在于，所述子矩阵运算单元包括子矩阵加法部件、子矩阵乘法部件、大小比较部件、非线性运算部件和子矩阵标量乘法部件，所述子矩阵加法部件、子矩阵乘法部件、大小比较部件、非线性运算部件和子矩阵标量乘法部件形成多流水级结构；

所述多流水级结构包括第一流水级、第二流水级和第三流水级，其中，所述子矩阵加法部件和子矩阵乘法部件处于第一流水级，大小比较部件处于第二流水级，非线性运算部件和子矩阵标量乘法部件处于第三流水级。

子矩阵运算装置及方法

[0001] 本申请是申请日为2016年04月22日、申请号为201610258546.6、专利名称为“一种子矩阵运算装置及方法”的分案申请。

技术领域

[0002] 本发明属于计算机领域，尤其涉及一种子矩阵运算装置及方法。

背景技术

[0003] 当前计算机领域有越来越多的算法涉及到矩阵运算，包括人工神经网络算法和图形的渲染算法。与此同时，作为矩阵运算中的一个重要组成部分，子矩阵运算也越来越频繁的出现在各种计算任务中。所以对于那些面向解决矩阵运算问题的方案，必须同时考虑子矩阵运算实现的效率和难度。

[0004] 在现有技术中一种进行子矩阵运算的已知方案是使用通用处理器，该方法通过通用寄存器堆和通用功能部件来执行通用指令，从而执行子矩阵运算。然而，该方法的缺点之一是单个通用处理器多用于标量计算，在进行子矩阵运算时运算性能较低。而使用多个通用处理器并行执行时，通用处理器之间的相互通讯又有可能成为性能瓶颈，同时，实现子矩阵运算的代码量也大于正常的矩阵运算。

[0005] 在另一种现有技术中，使用图形处理器(GPU)来进行子矩阵计算，其中，通过使用通用寄存器堆和通用流处理单元执行通用SIMD(Single Instruction Multiple Data，单指令多数据流)指令来进行子矩阵运算。然而，上述方案中，GPU片上缓存太小，在进行大规模子矩阵运算时需要不断进行片外数据搬运，片外带宽成为了主要性能瓶颈。

[0006] 在另一种现有技术中，使用专门定制的矩阵运算装置来进行子矩阵计算，其中，使用定制的寄存器堆和定制的处理单元进行子矩阵运算。然而，目前已有的专用矩阵运算装置受限于寄存器堆，子矩阵数据通常具有特定的规模，不能够灵活地支持不同长度的子矩阵运算。

[0007] 综上所述，现有的不管是片上多核通用处理器、片间互联通用处理器(单核或多核)、还是片间互联，图形处理器都无法进行高效的子矩阵运算，并且这些现有技术在处理子矩阵运算问题时存在着代码量大，受限于片间通讯，片上缓存不够，支持的子矩阵规模不够灵活等问题。

发明内容

[0008] 基于此，本发明提供一种子矩阵运算装置及方法，能配合子矩阵运算指令集，能够满足不同规模子矩阵数据的运算过程，高效地实现张量运算指令、子矩阵乘向量指令、向量乘子矩阵指令及子矩阵加减乘除等运算。

[0009] 一种子矩阵运算方法，所述方法包括：

[0010] 获取子矩阵运算指令，其中，所述子矩阵运算指令包括张量运算指令、子矩阵乘向量指令、向量乘子矩阵指令、子矩阵加法指令、子矩阵减法指令以及子矩阵乘法指令中的至

少一种；

[0011] 根据所述子矩阵运算指令分别从寄存器单元中获取第一子矩阵信息和第二子矩阵信息；

[0012] 根据所述第一子矩阵信息从存储单元中获取第一子矩阵数据，根据所述第二子矩阵信息从所述存储单元中获取第二子矩阵数据；

[0013] 根据所述第一子矩阵数据和所述第二子矩阵数据进行子矩阵运算，获得子矩阵运算结果。

[0014] 在其中一个实施例中，所述第一子矩阵信息包括所述第一子矩阵数据在所述存储单元中的起始地址、所述第一子矩阵数据的行宽、所述第一子矩阵数据的列宽以及所述第一子矩阵数据的行间隔，其中，所述第一子矩阵数据的行间隔是指所述第一子矩阵数据相邻两行间，上一行的行末数据到下一行的行首数据的数据间隔；所述第二子矩阵信息包括向量地址及向量长度；

[0015] 根据所述第一子矩阵数据和所述第二子矩阵数据进行子矩阵运算，获得子矩阵运算结果的步骤包括：

[0016] 将所述第一子矩阵数据作为被乘数，将所述第二子矩阵数据作为乘数进行子矩阵乘向量运算，获得子矩阵乘向量运算结果；

[0017] 或者，将所述第一子矩阵数据作为乘数，将所述第二子矩阵数据作为被乘数进行向量乘子矩阵运算，获得向量乘子矩阵运算结果。

[0018] 在其中一个实施例中，所述第一子矩阵信息和所述第二子矩阵信息分别包括对应子矩阵数据在所述存储单元中的起始地址、对应子矩阵数据的行宽、对应子矩阵数据的列宽以及对应子矩阵数据的行间隔，其中，子矩阵数据的行间隔是指所述子矩阵数据相邻两行间，上一行的行末数据到下一行的行首数据的数据间隔；

[0019] 根据所述第一子矩阵数据和所述第二子矩阵数据进行子矩阵运算，获得子矩阵运算结果的步骤包括：

[0020] 根据所述第一子矩阵数据和所述第二子矩阵数据进行矩阵加法运算或减法运算。

[0021] 在其中一个实施例中，所述第一子矩阵信息和所述第二子矩阵信息分别包括对应子矩阵数据在所述存储单元中的起始地址、对应子矩阵数据的行宽、对应子矩阵数据的列宽及对应子矩阵数据的行间隔，其中，所述子矩阵数据的行间隔是指所述子矩阵数据相邻两行间，上一行的行末数据到下一行的行首数据的数据间隔；

[0022] 根据所述第一子矩阵数据和所述第二子矩阵数据进行子矩阵运算，获得子矩阵运算结果的步骤包括：

[0023] 根据所述第一子矩阵数据和所述第二子矩阵数据进行对位乘法运算，获得子矩阵乘法运算结果。

[0024] 在其中一个实施例中，所述第一子矩阵信息和所述第二子矩阵信息分别包括对应子矩阵数据在所述存储单元中的起始地址、对应子矩阵数据的行宽、对应子矩阵数据的列宽及对应子矩阵数据的行间隔，其中，所述子矩阵数据的行间隔是指所述子矩阵数据相邻两行之间，上一行的行末数据到下一行的行首数据的数据间隔；

[0025] 根据所述第一子矩阵数据和所述第二子矩阵数据进行子矩阵运算，获得子矩阵运算结果的步骤包括：

[0026] 根据所述第一子矩阵数据和所述第二子矩阵数据进行张量运算，获得张量运算结果。

[0027] 在其中一个实施例中，所述子矩阵运算指令包括操作码和至少一个操作域，其中，所述操作码用于指示所述子矩阵运算指令的功能，操作域用于指示所述子矩阵运算指令的数据信息；

[0028] 所述子矩阵运算指令的数据信息包括所述寄存器单元的编号，从而能够根据寄存器单元的编号访问对应的寄存器单元，获取所述第一子矩阵信息和所述第二子矩阵信息。

[0029] 在其中一个实施例中，所述方法还包括如下步骤：

[0030] 对获取的子矩阵运算指令进行译码；

[0031] 判断所述子矩阵运算指令与前一子矩阵运算指令是否访问相同的子矩阵数据，若是，则将所述子矩阵运算指令存储在指令队列中，等待前一子矩阵运算指令执行完毕后，再执行根据所述子矩阵运算指令分别从寄存器单元中获取第一子矩阵信息和所述第二子矩阵信息步骤。

[0032] 本发明还提供了一种子矩阵运算装置，用于根据子矩阵运算指令从矩阵数据中获取子矩阵数据，并根据所述子矩阵数据执行子矩阵运算，所述装置包括：

[0033] 存储单元，用于存储矩阵数据；

[0034] 寄存器单元，用于存储子矩阵信息；

[0035] 子矩阵运算单元，用于获取子矩阵运算指令，根据所述子矩阵运算指令分别从所述寄存器单元中获取第一子矩阵信息和第二子矩阵信息；根据所述第一子矩阵信息从所述存储单元中获取第一子矩阵数据，根据所述第二子矩阵信息从所述存储单元中获取第二子矩阵数据；并根据所述第一子矩阵数据和所述第二子矩阵数据进行子矩阵运算，获得子矩阵运算结果；

[0036] 其中，所述子矩阵运算指令包括张量运算指令、子矩阵乘向量指令、向量乘子矩阵指令、子矩阵加法指令、子矩阵减法指令以及子矩阵乘法指令。

[0037] 在其中一个实施例中，所述子矩阵运算指令为子矩阵乘向量指令或向量乘子矩阵运算指令；所述第一子矩阵信息包括第一子矩阵数据在所述存储单元中的起始地址、第一子矩阵数据的行宽、第一子矩阵数据的列宽及第一子矩阵数据的行间隔，其中，所述第一子矩阵数据的行间隔是指第一子矩阵数据相邻两行间，上一行的行末数据到下一行的行首数据的数据间隔；所述第二子矩阵信息包括向量地址及向量长度。

[0038] 在其中一个实施例中，所述子矩阵运算指令为张量运算指令、子矩阵加法指令、子矩阵减法指令或子矩阵对位乘法指令；

[0039] 所述第一子矩阵信息和所述第二子矩阵信息分别包括对应子矩阵数据在所述存储单元中的起始地址、子矩阵数据的行宽、子矩阵数据的列宽、子矩阵数据的行间隔，其中，所述子矩阵数据的行间隔是指子矩阵数据相邻两行间，上一行的行末数据到下一行的行首数据的数据间隔。

[0040] 在其中一个实施例中，所述装置还包括用于获取所述子矩阵运算指令，并将所述子矩阵运算指令进行处理的指令处理单元；所述指令处理单元包括：

[0041] 取址模块，用于获取所述子矩阵运算指令；

[0042] 译码模块，用于对获取的所述子矩阵运算指令进行译码；

[0043] 指令队列，用于对译码后的所述子矩阵运算指令进行顺序存储；

[0044] 依赖关系处理单元，用于在所述子矩阵运算单元获取所述子矩阵运算指令之前，判断所述子矩阵运算指令与前一子矩阵运算指令是否访问相同的子矩阵数据，若是，则将所述子矩阵运算指令存储在指令队列中，等待前一子矩阵运算指令执行完毕后，再根据所述子矩阵运算指令分别获取第一子矩阵信息和第二子矩阵信息。

[0045] 在其中一个实施例中，所述存储单元还用于存储子矩阵运算结果；

[0046] 所述装置还包括输入输出单元，所述输入输出单元用于将矩阵数据存储至所述存储单元，所述输入输出单元还用于从所述存储单元中获取所述子矩阵运算结果。

[0047] 在其中一个实施例中，所述存储单元为高速暂存存储器。

[0048] 在其中一个实施例中，所述子矩阵运算单元包括子矩阵加法部件、子矩阵乘法部件、大小比较部件、非线性运算部件和子矩阵标量乘法部件，所述子矩阵加法部件、子矩阵乘法部件、大小比较部件、非线性运算部件和子矩阵标量乘法部件形成多流水级结构；

[0049] 所述多流水级结构包括第一流水级、第二流水级和第三流水级，其中，所述子矩阵加法部件和子矩阵乘法部件处于第一流水级，大小比较部件处于第二流水级，非线性运算部件和子矩阵标量乘法部件处于第三流水级。

[0050] 本发明提供的子矩阵运算方法及装置，可以根据子矩阵运算指令从寄存器单元中获取两个子矩阵信息，并分别根据两个子矩阵信息从存储单元中获取将参与子矩阵运算的两个子矩阵数据，且两个子矩阵数据可以具有不同的数据规模，然后可以根据获取的两个子矩阵数据进行子矩阵运算，获得子矩阵运算的结果，该子矩阵运算方法能够支持不同规模的子矩阵数据，提升包含大量矩阵计算任务的执行性能，同时提高了张量运算指令、子矩阵乘向量指令、向量乘子矩阵指令及子矩阵加减乘除等运算的运算效率。进一步地，本发明中的卷积指令能够支持不同矩阵长度，使用灵活方便。

附图说明

[0051] 图1是本申请一实施例提供的子矩阵运算装置的示意图；

[0052] 图2是本申请一实施例提供的指令集格式示意图；

[0053] 图3是本申请中一个子矩阵的示意图；

[0054] 图4是本申请另一实施例提供的子矩阵运算装置的示意图；

[0055] 图5是本申请实施例提供的子矩阵运算方法执行子矩阵乘子矩阵指令时的流程图；

[0056] 图6是本申请实施例中矩阵数据和子矩阵数据的示意图；

[0057] 图7是本申请实施例提供的子矩阵运算装置执行卷积神经网络运算的流程图。

具体实施方式

[0058] 本申请实施例提供了一种子矩阵运算装置及方法，包括存储单元、寄存器单元和子矩阵运算单元，存储单元中存储有子矩阵数据，寄存器单元中存储有子矩阵信息，子矩阵运算单元可以根据子矩阵运算指令在寄存器单元中获取子矩阵信息，然后，根据该子矩阵信息在存储单元中获取相应的子矩阵数据，接着，根据获取的子矩阵数据进行子矩阵运算，得到子矩阵运算结果。本申请实施例的存储单元可以为高速暂存存储器，通过将参与计算

的子矩阵数据暂存在高速暂存存储器上，使得子矩阵运算过程中可以更加灵活有效地支持不同宽度的数据，提升包含大量子矩阵计算任务的执行性能。其中高速暂存存储器可以通过各种不同存储器件，如静态RAM(SRAM)、动态RAM(DRAM)、增强动态RAM(EDRAM)、忆阻器、3D-DRAM和非易失存储等实现。

[0059] 图1是本申请实施例提供的子矩阵运算装置的示意图，如图1所示，该子矩阵运算装置包括存储单元、寄存器单元和子矩阵运算单元。其中，存储单元用于存储矩阵数据；寄存器单元用于存储子矩阵信息，在具体应用中，可以由多个寄存器单元组成一个寄存器堆，每个寄存器单元存储有不同的子矩阵信息，需要说明书的是，子矩阵信息均为标量数据。可选地，子矩阵信息可以包括子矩阵数据在存储单元中的起始地址(start_addr)、子矩阵数据的行宽(iter1)、子矩阵数据的列宽(iter2)以及行间隔(stride1)，其中，行间隔是指子矩阵数据相邻两行间，上一行的行末数据到下一行的行首数据的数据间隔。

[0060] 如图3所示，矩阵数据实际在存储单元中是以一维的方式存储的，子矩阵的起始地址即图3中子矩阵左上角元素的地址，子矩阵的行宽即图3中子矩阵每一行元素的个数，子矩阵的列宽即图3中子矩阵每一列元素的个数，子矩阵的行间距即图3中子矩阵上一行最后一个元素到下一行第一个元素之间的地址间距。该子矩阵运算装置在实际读取子矩阵数据时，只需要从子矩阵数据在存储单元中的起始位置start_addr开始，每读取iter1个数据后跳过stride1个数据再读取iter1个数据，重复iter2次即可获得完整的子矩阵数据。这样，通过上述方式获得的子矩阵数据可以是规模不定的矩阵数据，即子矩阵数据的行宽、列宽以及行间隔中的一个或多个可以是不固定的。相对于现有技术中，子矩阵规模固定的运算装置，本申请实施例的装置获取的子矩阵运算装置，能够支持不同规模的子矩阵数据，提升了包含大量矩阵计算任务的执行性能。

[0061] 子矩阵运算单元用于获取子矩阵运算指令，并根据该子矩阵运算指令从寄存器单元中获取子矩阵信息，然后，根据该子矩阵信息从存储单元中的矩阵数据中获取子矩阵数据，接着，根据获取的子矩阵数据进行子矩阵运算，得到子矩阵运算结果。可选地，该子矩阵运算可以包括卷积运算、张量运算、子矩阵乘向量运算、向量乘子矩阵运算、子矩阵对位乘法运算、子矩阵加法运算及子矩阵减法运算等以及子矩阵搬运运算等等。本申请实施例中，每个子矩阵运算可以通过子矩阵运算指令实现，且子矩阵运算指令具有特定的指令格式。

[0062] 图2是本申请实施例提供的指令集格式示意图，如图2所示，指令集采用Load/Store结构，子矩阵运算单元不会对内存中的数据进行操作。子矩阵指令集采用超长指令集架构(Very Long Instruction Word)，同时，指令集采用定长指令，使得子矩阵运算装置在上一条子矩阵运算指令的译码阶段就可以对下一条子矩阵运算指令进行取值。可选地，子矩阵运算指令可以包括操作码和多个操作域，其中，操作码用于指示该子矩阵运算指令的功能，操作域用于指示该子矩阵运算指令的数据信息，数据信息为寄存器单元的编号或者立即数，子矩阵运算单元可以根据寄存器单元的编号访问对应的寄存器单元，从而获取子矩阵信息。或者，子矩阵运算单元也可以直接将立即数作为子矩阵数据进行相应的子矩阵运算。

[0063] 需要说明的是，针对不同功能的运算指令，子矩阵运算指令的操作码也不同，具体地，在本申请实施例提供的一套指令集中，包含有不同功能的子矩阵运算指令：

[0064] 子矩阵乘向量指令(SMMV)，根据该指令，装置从高速暂存存储器的指定起始地址，

根据指令中子矩阵的行宽、列宽和行间距取出指定的子矩阵数据，同时取出向量数据，在子矩阵运算单元中进行矩阵乘向量的乘法运算，并将结果写回至高速暂存存储器的指定地址；值得说明的是，向量可以作为特殊形式的矩阵（只有一行元素的矩阵）存储于高速暂存存储器中。

[0065] 向量乘子矩阵指令（VMSM），根据该指令，装置从高速暂存存储器的指定地址取出向量数据，同时根据指令中的子矩阵起始地址、子矩阵的行宽和列宽以及子矩阵的行间距取出指定的子矩阵，在矩阵单元中进行向量乘子矩阵的乘法运算，并将结果写回至高速暂存存储器的指定地址；值得说明的是，向量可以作为特殊形式的矩阵（只有一行元素的矩阵）存储于高速暂存存储器中。

[0066] 子矩阵乘标量指令（SMMS），根据该指令，装置从高速暂存存储器的指定地址，根据指令中的子矩阵的行宽和列宽以及子矩阵的行间距，取出指定的子矩阵数据，从标量寄存器堆的指定地址中取出指定的标量数据，在子矩阵运算单元中进行子矩阵乘标量的运算，并将结果写回至高速暂存存储器的指定地址，需要说明的是，标量寄存器堆不仅存储有子矩阵的各种数据信息（包括起始地址、行宽、列宽和行间距），还存有标量数据本身。

[0067] 张量运算指令（TENS），根据该指令，装置从高速暂存存储器取出指定的两块子矩阵数据，在子矩阵运算单元中对两子矩阵数据进行张量运算，并将计算结果写回至高速暂存存储器的指定地址。本领域技术人员可以理解的是，在一个坐标系下，张量是由若干个分量来表示，而在不同坐标系下的分量之间应满足一定的变换规则，如矩阵、多变量线性形式等。张量可以包括一阶张量、二阶张量及m阶张量（m表示张量的维度），各阶张量均可以采用矩阵进行表示。例如，一阶张量又称矢量或向量，可以采用1×n的行向量表示，其中，n表示向量的长度；二阶张量是有m^2个数组组成，其中，m表示张量的维度。张量的基本运算可以包括张量的加减运算、张量的乘法运算及张量函数的求导运算等等。

[0068] 子矩阵加法指令（SMA），根据该指令，装置从高速暂存存储器取出指定的两块子矩阵数据，在子矩阵运算单元中对两子矩阵数据进行加法运算，并将计算结果写回至高速暂存存储器的指定地址。

[0069] 子矩阵减法指令（SMS），根据该指令，装置从高速暂存存储器取出指定的两块子矩阵数据，在子矩阵运算单元中对两子矩阵数据进行减法运算，并将计算结果写回至高速暂存存储器的指定地址。

[0070] 子矩阵乘法指令（SMM），根据该指令，装置从高速暂存存储器取出指定的两块子矩阵数据，在子矩阵运算单元中对两子矩阵数据进行对位乘法运算，并将计算结果写回至高速暂存存储器的指定地址。本领域技术人员可以理解的是，也可以通过子矩阵乘法指令实现子矩阵对位相除运算，因此，该子矩阵运算装置还可以执行子矩阵除法运算。

[0071] 卷积指令（CONV），根据该指令，实现用卷积核对矩阵进行卷积滤波。装置从高速暂存存储器取出指定的卷积核矩阵，从待卷积矩阵存储的起始地址开始，对当前位置下卷积核覆盖的子矩阵数据进行滤波，即在子矩阵运算单元中对卷积核和子矩阵进行对位乘法运算，并对得到的矩阵进行元素求和，得到当前位置的滤波结果，将结果写回至高速暂存存储器的指定地址。然后根据指令中给定的位移参数，在待卷积矩阵上移动至下一位置，重复上面的运算，直到移动至结束位置。

[0072] 子矩阵搬运指令（SMMOVE），根据该指令，装置将高速暂存存储器中存储的指定子

矩阵存至高速暂存存储器的另一处地址。

[0073] 进一步，子矩阵运算装置还包括指令处理单元，用于获取子矩阵运算指令，并对该子矩阵运算指令进行处理后，提供给子矩阵运算单元。具体地，如图4所示，指令处理单元可以包括取指模块、译码模块、指令队列及依赖关系处理单元，其中，取指模块用于获取子矩阵运算指令，译码模块用于对获取的子矩阵运算指令进行译码，指令队列用于对译码后的子矩阵运算指令进行顺序存储，依赖关系处理单元用于在子矩阵运算单元获取子矩阵运算指令前，判断该子矩阵运算指令与前一子矩阵运算指令是否访问相同的子矩阵数据，若是，则将该子矩阵运算指令存储在指令队列中，等待前一子矩阵运算指令执行完毕后，再将所述指令队列中的该子矩阵运算指令提供给所述子矩阵运算单元，否则，直接将该子矩阵运算指令提供给所述子矩阵运算单元。

[0074] 进一步，存储单元还用于存储子矩阵运算结果，优选地，可采用高速暂存存储器作为存储单元。另外，本发明还包括输入输出单元，其与存储单元直接连接，输入输出单元用于将矩阵数据存储于存储单元，或者，从存储单元中获取子矩阵运算结果。

[0075] 进一步，子矩阵运算单元还可以包括子矩阵加法部件、子矩阵乘法部件、大小比较部件、非线性运算部件和子矩阵标量乘法部件。进一步地，子矩阵运算单元为多流水级结构，多流水级结构包括第一流水级、第二流水级和第三流水级，其中，子矩阵加法部件和子矩阵乘法部件处于第一流水级，大小比较部件处于第二流水级，非线性运算部件和子矩阵标量乘法部件处于第三流水级。

[0076] 本申请实施例还提供一种子矩阵运算方法，包括：

[0077] S1，存储矩阵数据；

[0078] S2，存储子矩阵信息；

[0079] S3，获取子矩阵运算指令，并根据该子矩阵运算指令获取子矩阵信息，然后，根据该子矩阵信息从存储的矩阵数据中获取子矩阵数据，接着，根据获取的子矩阵数据进行子矩阵运算，得到子矩阵运算结果。

[0080] 进一步，在步骤S3之前，还包括：

[0081] 获取子矩阵运算指令；

[0082] 对获取的子矩阵运算指令进行译码；

[0083] 判断该子矩阵运算指令与前一子矩阵运算指令是否访问相同的子矩阵数据，若是，则将该子矩阵运算指令存储在指令队列中，等待前一子矩阵运算指令执行完毕后，再将执行所述步骤S3，否则，直接执行步骤S3。

[0084] 进一步，步骤S3还包括，存储子矩阵运算结果。

[0085] 进一步，上述方法还包括：步骤S4，获取存储的子矩阵运算结果。

[0086] 进一步，子矩阵运算包括子矩阵加法运算、子矩阵乘法运算、大小比较运算、非线性运算和子矩阵标量乘法运算。进一步地，采用多流水级结构进行子矩阵运算，多流水级结构包括第一流水级、第二流水级和第三流水级，其中，在第一流水级进行子矩阵加法运算和子矩阵乘法运算，在第二流水级进行大小比较运算，在第三流水级进行非线性运算和子矩阵标量乘法运算。

[0087] 例如，图4是本发明实施例提供的子矩阵运算装置的示意图，如图4所示，装置包括取指模块、译码模块、指令队列、标量寄存器堆(即寄存器单元)、依赖关系处理单元、指令队

列、子矩阵运算单元、高速暂存器(即存储单元)、IO内存存取模块(即输入输出单元),其中:

[0088] 取指模块用于从指令序列中取出下一条将要执行的指令,并将该指令传给译码模块;

[0089] 译码模块用于对获取的指令进行译码,并将译码后指令传给指令队列;

[0090] 考虑到不同指令在包含的标量寄存器上有可能存在依赖关系,指令队列用于缓存译码后的指令,当依赖关系被满足之后发送指令;

[0091] 标量寄存器堆能够提供装置在运算过程中所需的多个标量寄存器;

[0092] 依赖关系处理单元用于处理指令与前一条指令可能存在的存储依赖关系。子矩阵运算指令会访问高速暂存存储器,前后指令可能会访问同一块存储空间。为了保证指令执行结果的正确性,该指令如果被检测到与之前的指令的数据存在依赖关系,该指令必须在指令队列内等待至依赖关系被消除。

[0093] 指令队列是一个有序队列,与之前指令在数据上有依赖关系的指令被存储在该队列内直至存储关系被消除;

[0094] 子矩阵运算单元,该模块负责装置的所有子矩阵运算,包括但不限于子矩阵加法操作、子矩阵加标量操作、子矩阵减法操作、子矩阵减标量操作、子矩阵乘法操作、子矩阵乘标量操作、子矩阵除法(对位相除)操作、子矩阵与操作和子矩阵或操作,子矩阵运算指令被送往该运算单元执行。

[0095] 高速暂存存储器,该模块是矩阵数据专用的暂存存储装置,能够支持不同大小的矩阵数据;

[0096] IO内存存取模块,该模块用于直接访问高速暂存存储器,负责从高速暂存存储器中读取数据或写入数据。

[0097] 本领域的技术人员根据上文的记载可以毫无疑义地获知,当上述子矩阵运算装置在执行张量运算指令、子矩阵乘向量指令、向量乘子矩阵指令、子矩阵加法指令、子矩阵减法指令以及子矩阵乘法指令等指令,需要从存储单元中获得两块子矩阵数据,并根据获取的两块子矩阵数据进行子矩阵运算。

[0098] 在一个实施例中,上述子矩阵运算方法可以包括如下步骤:

[0099] 获取子矩阵运算指令,具体地,子矩阵运算单元可以获取子矩阵运算指令。其中,子矩阵运算指令包括张量运算指令、子矩阵乘向量指令、向量乘子矩阵指令、子矩阵加法指令、子矩阵减法指令以及子矩阵乘法指令。当然,在其他实施例中,该子矩阵运算指令还可以包括卷积指令、子矩阵搬运指令及子矩阵乘标量指令等等。进一步地,还可以对获取的子矩阵运算指令进行指令预处理操作。即上述方法还可以包括如下步骤:译码模块对获取的子矩阵运算指令进行译码;依赖关系处理单元判断获取的子矩阵运算指令与前一子矩阵运算指令是否访问相同的子矩阵数据,若是,则将子张量运算指令存储在指令队列中,等待前一子矩阵运算指令执行完毕后,再执行根据张量运算指令分别获取第一子矩阵信息和第二子矩阵信息步骤。

[0100] 根据子矩阵运算指令分别从寄存器单元中获取两个子矩阵信息,两个子矩阵信息可以分别表示为第一子矩阵信息和第二子矩阵信息。具体地,子矩阵运算单元可以根据子矩阵运算指令分别从寄存器单元中获取两个子矩阵信息,两个子矩阵信息可以分别表示为第一子矩阵信息和第二子矩阵信息。可选地,第一子矩阵信息和第二子矩阵信息可以包括

相应的子矩阵数据在存储单元中的起始地址(start_addr)、子矩阵数据的行宽(iter1)、子矩阵数据的列宽(iter2)以及行间隔(stride1)，其中，行间隔是指子矩阵数据相邻两行间，上一行的行末数据到下一行的行首数据的数据间隔。当然，第一子矩阵信息或第二子矩阵信息还可以仅包括相应子矩阵数据在存储单元中的向量地址及向量长度，该向量地址可以是子矩阵数据在存储单元中的起始地址、向量长度可以包括子矩阵数据的行宽及子矩阵数据的列宽，其中，行宽或列宽的取值可以为1。

[0101] 根据第一子矩阵信息从存储单元中获取第一子矩阵数据，根据第二子矩阵信息从存储单元中获取第二子矩阵数据；具体地，子矩阵运算单元可以根据第一子矩阵信息从存储单元中获取第一子矩阵数据，根据第二子矩阵信息从存储单元中获取第二子矩阵数据。本申请实施例中，根据第一子矩阵信息获取第一子矩阵数据的过程可参见图3及上文的描述，根据第二子矩阵信息获取第二子矩阵数据的过程也可参见图3及上文的描述。

[0102] 根据第一子矩阵数据和第二子矩阵数据进行子矩阵运算，获得子矩阵运算结果。具体地，子矩阵运算单元可以根据第一子矩阵数据和第二子矩阵数据进行子矩阵运算，获得子矩阵运算结果。该子矩阵运算可以包括张量运算、子矩阵乘向量运算、向量乘子矩阵运算、子矩阵加减运算、子矩阵对位乘法运算(子矩阵对位除法运算)及卷积运算等运算。

[0103] 可选地，该子矩阵指令可以为子矩阵乘向量指令或向量乘子矩阵指令。此时，第一子矩阵信息包括第一子矩阵数据在存储单元中的起始地址、第一子矩阵数据的行宽、第一子矩阵数据的列宽以及第一子矩阵数据的行间隔，其中，第一子矩阵数据的行间隔是指第一子矩阵数据相邻两行间，上一行的行末数据到下一行的行首数据的数据间隔；第二子矩阵信息包括向量地址及向量长度；

[0104] 子矩阵运算单元根据第一子矩阵数据和第二子矩阵数据进行子矩阵运算，获得子矩阵运算结果的步骤包括：

[0105] 子矩阵运算单元将第一子矩阵数据作为被乘数，将第二子矩阵数据作为乘数进行子矩阵乘向量运算，获得子矩阵乘向量运算结果；

[0106] 或者，子矩阵运算单元将第一子矩阵数据作为乘数，将第二子矩阵数据作为被乘数进行向量乘子矩阵运算，获得向量乘子矩阵运算结果。

[0107] 例如，图5是本发明实施例提供的子矩阵运算装置执行子矩阵乘向量执行的流程图，如图5所示，执行子矩阵乘向量指令的过程包括：

[0108] S1，取指模块取出该条子矩阵乘向量指令，并将该指令送往译码模块。

[0109] S2，译码模块对指令译码，并将指令送往指令队列。

[0110] S3，在指令队列中，该子矩阵乘向量指令需要从标量寄存器堆中获取指令中操作域所对应的标量寄存器里的数据，包括输入向量地址、输入向量长度、输入子矩阵地址、输入子矩阵行宽、输入子矩阵列宽、输入子矩阵行间距、输出向量地址、输出向量长度。

[0111] S4，在取得需要的标量数据后，该指令被送往依赖关系处理单元。依赖关系处理单元分析该指令与前面的尚未执行结束的指令在数据上是否存在依赖关系。该条指令需要在指令队列中等待至其与前面的未执行结束的指令在数据上不再存在依赖关系为止。

[0112] S5，依赖关系不存在后，该条子矩阵乘向量指令被送往矩子阵运算单元。子矩阵运算单元根据所需数据的位置信息从高速暂存器中取出需要的子矩阵和向量数据，然后在子矩阵运算单元中完成乘法运算。

[0113] S6,运算完成后,将结果写回至高速暂存存储器的指定地址。

[0114] 本领域的技术人员可以理解的是,向量乘子矩阵指令的执行过程与上述子矩阵乘向量指令的流程类似,其不同之处仅在于乘数与被乘数的位置变化。

[0115] 可选地,第一子矩阵信息和第二子矩阵信息分别包括对应子矩阵数据在存储单元中的起始地址、对应子矩阵数据的行宽、对应子矩阵数据的列宽以及对应子矩阵数据的行间隔,其中,子矩阵数据的行间隔是指对应的子矩阵数据相邻两行间,上一行的行末数据到下一行的行首数据的数据间隔。具体地,第一子矩阵信息可以包括第一子矩阵数据在存储单元中的起始地址、第一子矩阵数据的行宽、第一子矩阵数据的列宽以及行间隔等。第二子矩阵信息可以包括第二子矩阵数据在存储单元中的起始地址、第二子矩阵数据的行宽、第二子矩阵数据的列宽以及行间隔等。此时,子矩阵运算指令可以是张量运算指令、子矩阵加法指令、子矩阵减法指令以及子矩阵乘法指令。

[0116] 若子矩阵运算指令是子矩阵加减运算指令,此时,子矩阵运算单元根据第一子矩阵数据和第二子矩阵数据进行子矩阵运算的步骤还可以包括:

[0117] 子矩阵运算单元根据第一子矩阵数据和第二子矩阵数据进行矩阵加法运算或减法运算。

[0118] 若子矩阵运算指令是子矩阵乘法指令或子矩阵除法指令,此时子矩阵运算单元根据第一子矩阵数据和第二子矩阵数据进行子矩阵运算的步骤还可以包括:

[0119] 子矩阵运算单元根据第一子矩阵数据和第二子矩阵数据进行对位乘法运算,获得子矩阵乘法运算结果。

[0120] 若子矩阵运算指令为张量运算指令,此时,子矩阵运算单元根据第一子矩阵数据和第二子矩阵数据进行子矩阵运算的步骤还可以包括:

[0121] 子矩阵运算单元根据第一子矩阵数据和第二子矩阵数据进行张量运算,获得张量运算结果。本领域技术人员可以理解的是,张量的基本运算可以包括张量的加减运算、张量的乘法运算及张量函数的求导运算等等。

[0122] 可选地,该子矩阵运算方法还可以用于根据卷积运算指令从待卷积矩阵数据中获取子矩阵数据,并根据子矩阵数据执行卷积运算。具体地,上述子矩阵运算方法可以包括如下步骤:

[0123] 获取卷积指令;具体地,子矩阵运算单元可以获取卷积指令。进一步地,指令处理单元的取指模块可以读取卷积指令,指令处理单元的译码模块可以对获取的卷积指令进行译码,指令处理单元的依赖关系处理单元可以判断该卷积指令与前一子矩阵运算指令是否访问相同的子矩阵数据,若是,则将该卷积指令存储在指令队列中,等待前一子矩阵运算指令执行完毕,之后,指令处理单元可以将该卷积指令传送至子矩阵运算单元。

[0124] 根据卷积指令从存储单元中获取卷积核矩阵数据;本申请实施例中,可以通过IO指令将待卷积的矩阵数据和卷积核矩阵数据存储在存储单元的指定地址。当子矩阵运算单元获取到卷积指令后,其可以根据该卷积指令从存储单元中获取卷积核矩阵数据。

[0125] 从待卷积矩阵的起始位置开始,获取卷积核矩阵数据在当前位置的子矩阵数据;可选地,子矩阵运算单元可以从待卷积矩阵的起始位置开始,根据卷积指令从寄存器单元中获取所述卷积核矩阵数据在所述当前位置对应的子矩阵信息,之后,子矩阵运算单元可以根据当前位置对应的子矩阵信息从存储单元中获取卷积核矩阵数据在当前位置的子矩

阵数据。其中，子矩阵信息包括子矩阵数据在存储单元中的起始地址、子矩阵数据的行宽、子矩阵数据的列宽、以及行间隔，其中，行间隔是指子矩阵数据相邻两行间，上一行的行末数据到下一行的行首数据的数据间隔。

[0126] 执行卷积计算操作，该卷积计算操作包括：对卷积核矩阵数据和卷积核矩阵数据在当前位置的子矩阵数据进行对位相乘运算获得多个元素，并对多个元素进行累加求和运算，获得当前位置的卷积结果。也就是说，本申请实施例中，子矩阵运算单元采用对位相乘求和法进行卷积运算。

[0127] 根据卷积指令中给定的位移参数，将卷积核矩阵数据从当前位置移动至下一位置，并获取下一位置对应的子矩阵数据，之后返回执行卷积计算操作的步骤，直至完成待卷积矩阵数据的卷积计算，获得结果矩阵。子矩阵运算单元可以重复上述的位移步骤及卷积计算操作，直至获得结果矩阵，之后，可以将该结果矩阵存储至片外。

[0128] 具体地，图6为本发明实施例提供的子矩阵运算单元进行卷积神经网络运算的方法的流程图，该方法主要由子矩阵运算指令实现。卷积神经网络的运算特征是：对于n×y×x规模的特征图像输入(其中n是输入特征图像数，y是特征图像长，x是特征图像宽)，有n×h×w规模的卷积核，卷积核在输入图像上不断移动，在每个位置卷积核与自己所覆盖的输入图像的数据进行卷积运算，得到输出图像上对应的一个点的值。针对这种运算特征，卷积神经网络可以由一条子矩阵卷积指令循环实现。在实际的存储中，如图6所示，数据存储时在图像个数的维度上展开，输入数据图像由n×y×x的三维数组变成y×(x×n)的二维矩阵，相同地，卷积核数据变成h×(w×n)的二维矩阵。如图7所示，实现卷积神经网络的过程包括：

[0129] S1，通过IO指令将待卷积的矩阵数据和卷积核矩阵数据存至矩阵专用高速暂存存储器的指定地址；

[0130] S2，译码器取出CONV运算指令，根据该指令，子矩阵运算单元从高速暂存存储器中读取卷积核矩阵数据和该卷积核在输入图像起始位置的子矩阵数据。

[0131] S3，两矩阵数据在子矩阵运算单元中进行对位相乘和元素累加求和的运算，并写回结果。然后子矩阵运算单元继续读入卷积核，同时根据指令中位移参数得到的下一个待卷积的子矩阵的起始地址，读取数据。

[0132] S4，在CONV指令执行过程中，上面过程不断循环，直到完成卷积核在待卷积矩阵最后一个位置上的卷积运算。

[0133] S5，通过IO指令将卷积后的结果矩阵存至片外。

[0134] 需声明，本实施例采用了一种更加高效的方法实现卷积运算，即将三维的输入图像和卷积核均展开成二维形式，实际上，这不是本发明的装置和方法实现卷积运算的唯一方式，一种更通用的方法是对输入的每一张二维图像，与对应的卷积核中的一个面通过子矩阵指令执行卷积运算，得到输出结果的一个部分和，最终的卷积结果是所有的二维图像和与之相对应的卷积核中的面进行卷积运算得到的部分和的累加。故，子矩阵运算指令可以以多种方式实现卷积操作。

[0135] 综上所述，本申请提供子矩阵运算装置，并配合相应的子矩阵运算指令集，能够很好地解决当前计算机领域越来越多的算法包含大量子矩阵运算的问题，相比于已有的传统解决方案，本申请可以使用方便、支持的子矩阵规模灵活、片上缓存充足等优点。本发明可

以用于多种包含大量子矩阵运算的计算任务，包括目前表现十分出色的人工神经网络算法的反向训练和正向预测。

[0136] 以上实施例的各技术特征可以进行任意的组合，为使描述简洁，未对上述实施例中的各个技术特征所有可能的组合都进行描述，然而，只要这些技术特征的组合不存在矛盾，都应当认为是本说明书记载的范围。

[0137] 以上所述实施例仅表达了本申请的几种实施方式，其描述较为具体和详细，但并不能因此而理解为对发明专利范围的限制。应当指出的是，对于本领域的普通技术人员来说，在不脱离本申请构思的前提下，还可以做出若干变形和改进，这些都属于本申请的保护范围。因此，本申请专利的保护范围应以所附权利要求为准。

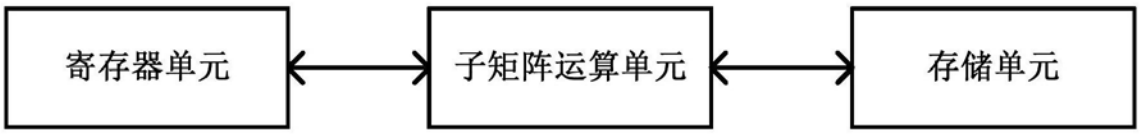


图1

操作码	寄存器或立即数	寄存器/立即数	...

图2

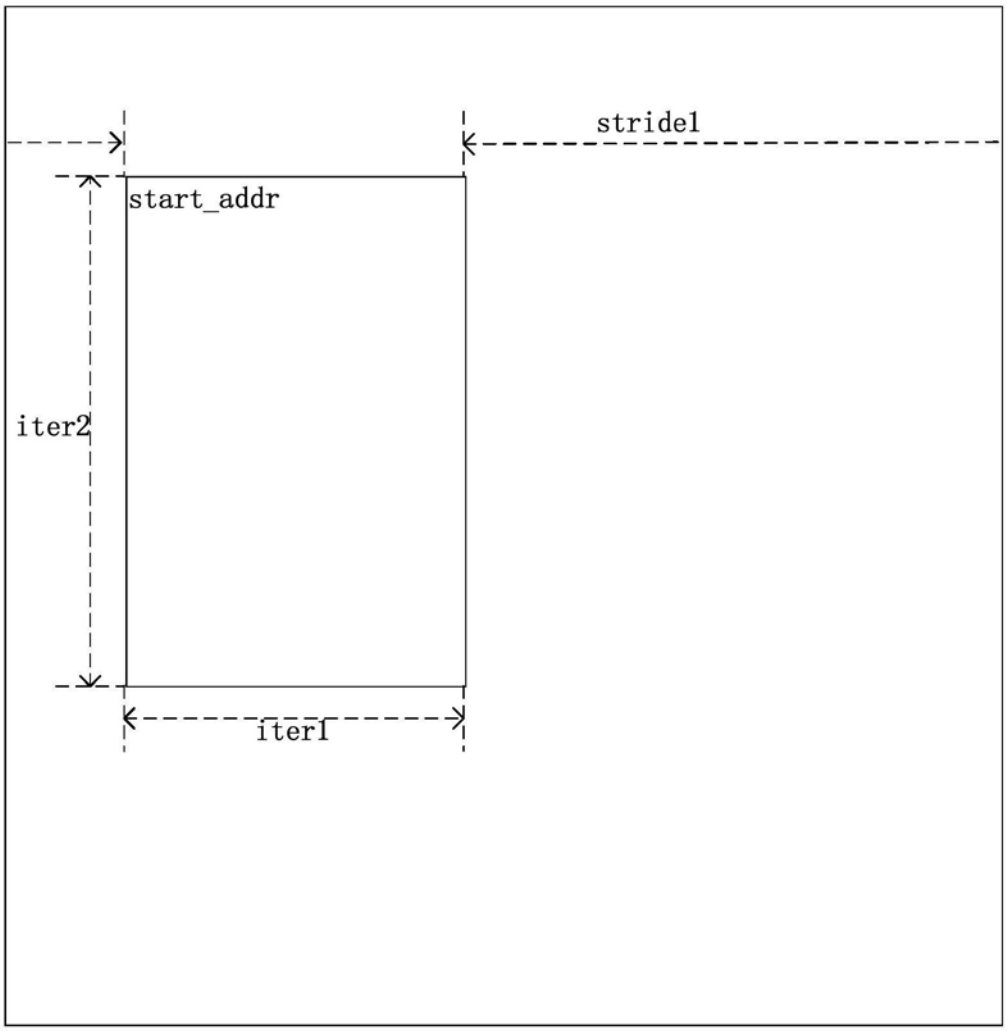


图3

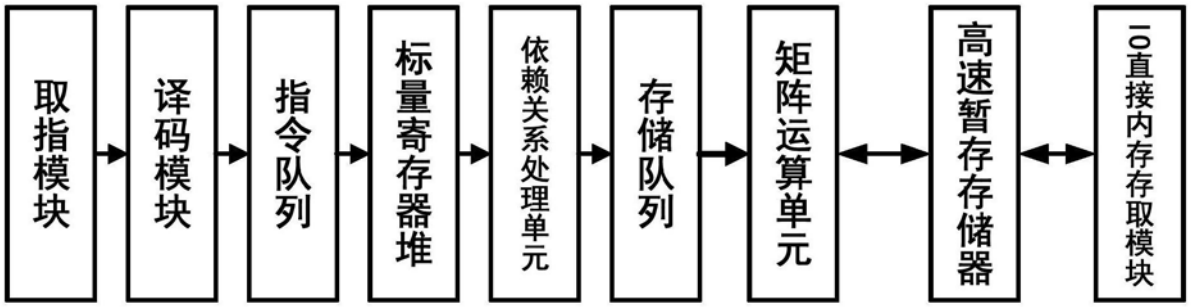

取指模块
译码模块
指令队列
标量寄存器堆
依赖关系处理单元
存储队列
矩阵运算单元
高速暂存存储器
IO直接内存存取模块

图4

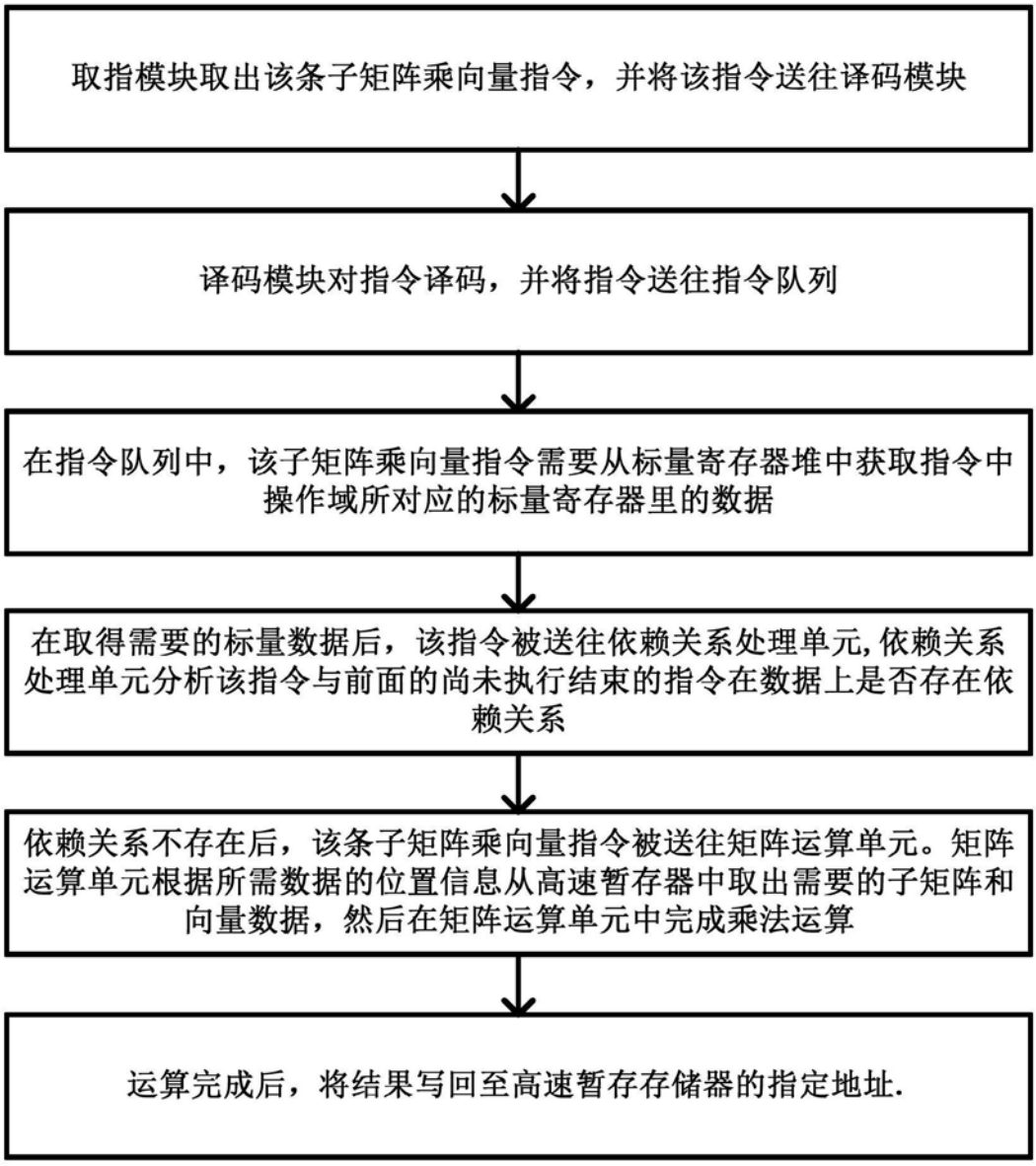

取指模块取出该条子矩阵乘向量指令，并将该指令送往译码模块
译码模块对指令译码，并将指令送往指令队列
在指令队列中，该子矩阵乘向量指令需要从标量寄存器堆中获取指令中操作域所对应的标量寄存器里的数据
在取得需要的标量数据后，该指令被送往依赖关系处理单元,依赖关系处理单元分析该指令与前面的尚未执行结束的指令在数据上是否存在依赖关系
依赖关系不存在后，该条子矩阵乘向量指令被送往矩阵运算单元。矩阵运算单元根据所需数据的位置信息从高速暂存器中取出需要的子矩阵和向量数据，然后在矩阵运算单元中完成乘法运算
运算完成后，将结果写回至高速暂存存储器的指定地址.

图5

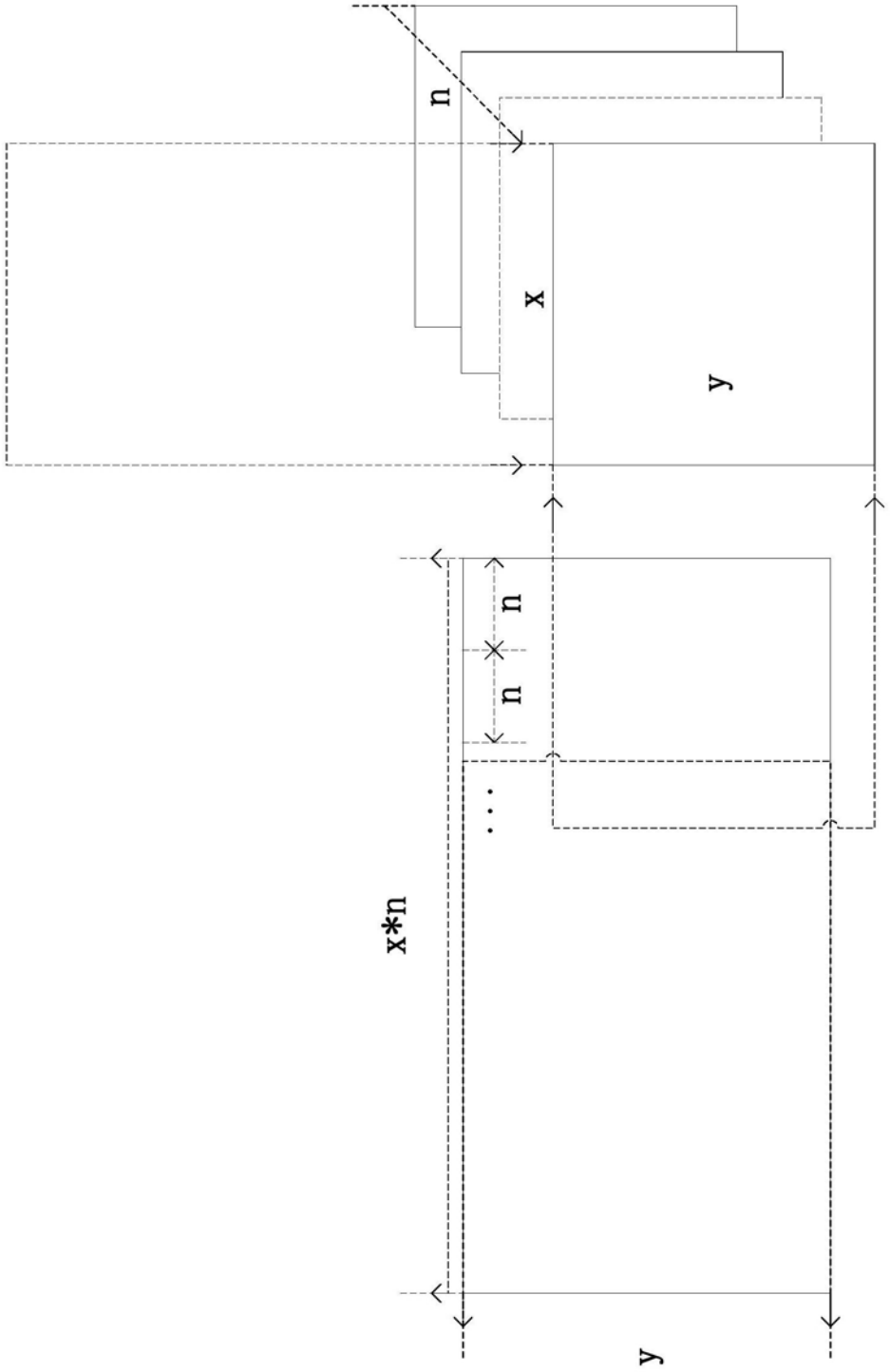


图6

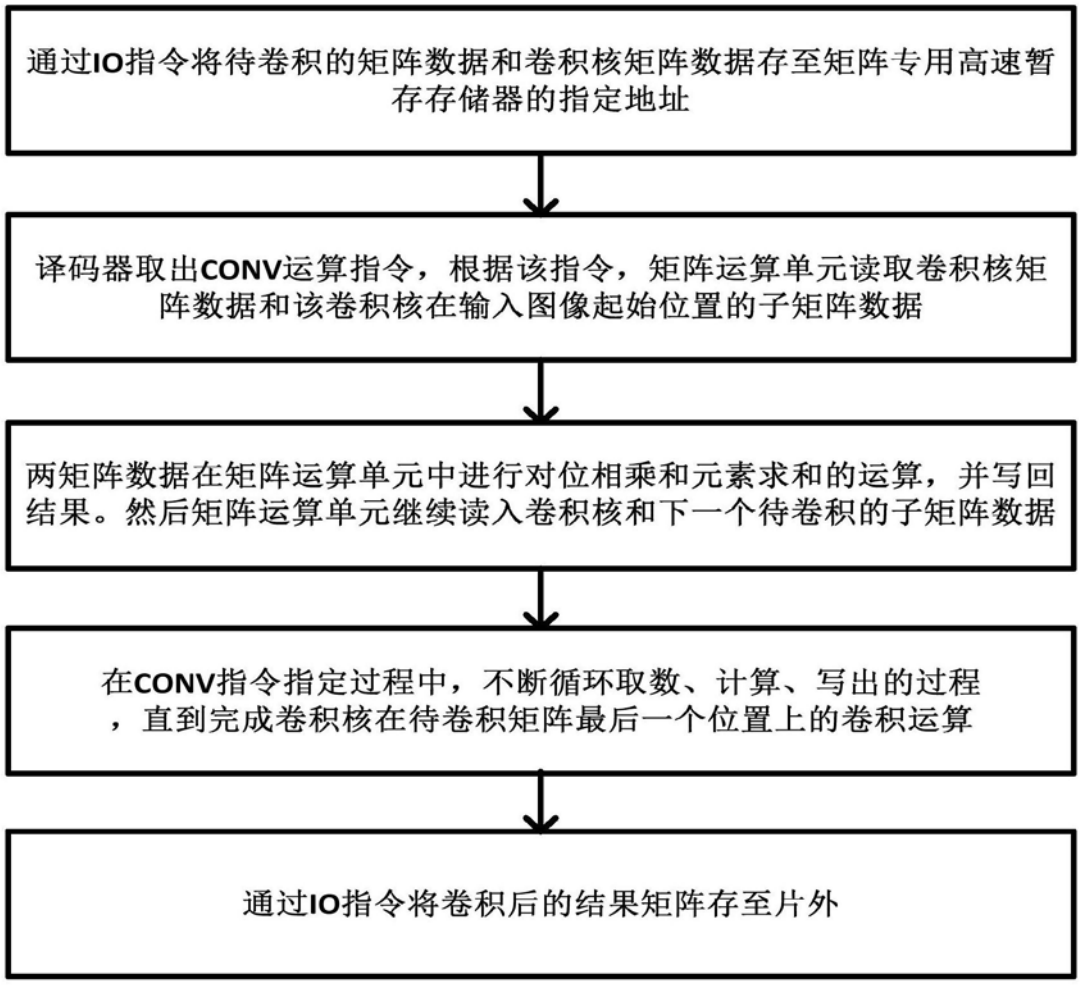

通过IO指令将待卷积的矩阵数据和卷积核矩阵数据存至矩阵专用高速暂存存储器的指定地址
译码器取出CONV运算指令，根据该指令，矩阵运算单元读取卷积核矩阵数据和该卷积核在输入图像起始位置的子矩阵数据
两矩阵数据在矩阵运算单元中进行对位相乘和元素求和的运算，并写回结果。然后矩阵运算单元继续读入卷积核和下一个待卷积的子矩阵数据
在CONV指令指定过程中，不断循环取数、计算、写出的过程，直到完成卷积核在待卷积矩阵最后一个位置上的卷积运算
通过IO指令将卷积后的结果矩阵存至片外

图7