(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113077042 B
(45)授权公告日 2024.06.04

(21)申请号 202011594777.7

(22)申请日 2020.12.29

(65)同一申请的已公布的文献号
申请公布号 CN 113077042 A

(43)申请公布日 2021.07.06

(30)优先权数据
16/734,792 2020.01.06 US

(73)专利权人 平头哥(上海)半导体技术有限公司
地址 200120 上海市浦东新区自由贸易试验区上科路366号、川和路55弄2号5层

(72)发明人 焦阳 陈龙 苏奕荣

(74)专利代理机构 北京清源汇知识产权代理事务所(特殊普通合伙) 11644
专利代理师 冯德魁

(51)Int.Cl.
G06N 3/0464(2023.01)
G06N 3/063(2023.01)

(56)对比文件
CN 107679621 A,2018.02.09
CN 107844826 A,2018.03.27
CN 108734636 A,2018.11.02
CN 109993277 A,2019.07.09
CN 110337807 A,2019.10.15
CN 114761925 A,2022.07.15
US 2018307980 A1,2018.10.25
张军阳;郭阳.二维矩阵卷积在向量处理器中的设计与实现.国防科技大学学报.2018,(03),全文.
周国昌;张立新.基于RCSIMD的8192点FFT并行算法研究.微电子学与计算机.2011,(04),全文.

审查员 曹宁

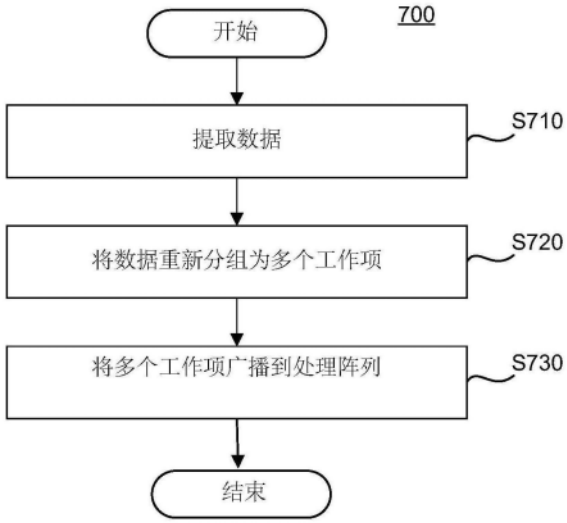
权利要求书3页 说明书16页 附图11页

(54)发明名称

卷积神经网络的数据重用与高效处理方法

(57)摘要

本发明涉及一种用于执行卷积神经网络操作的装置。该装置包括第一内存;包含多个处理字符串的处理阵列、以及控制器,该控制器能被配置为:将一个或多个批次的数据提取到所述第一内存中,将所述一个或多个批次的数据重新分组为多个工作项,其中,第一工作项部分地重叠于所述多个工作项中的一个或多个工作项,以及将所述多个工作项广播到所述处理阵列,其中,将所述第一工作项传送到所述处理阵列的两个或多个处理字符串。



CN 113077042 B

1.一种用于执行卷积神经网络操作的装置,包括:

第一内存;

包含多个处理字符串的处理阵列;以及

控制器,被配置为:

将一个或多个批次的数据提取到所述第一内存中;

将所述一个或多个批次的数据重新分组为多个工作项,其中,第一工作项部分地重叠于所述多个工作项中的一个或多个工作项;以及

将所述多个工作项广播到所述处理阵列,其中,将所述第一工作项传送到所述处理阵列的两个或多个处理字符串。

2.如权利要求1所述的装置,其中,将所述多个处理字符串分类为多个子集,并且将所述第一工作项传送到所述多个子集中的每个子集的第一处理字符串。

3.如权利要求2所述的装置,进一步包括第二存储器,所述第二存储器存储多个滤波器,所述多个滤波器的数目对应于所述子集的数目。

4.如权利要求1所述的装置,其中,每个所述处理字符串包括一个乘法器和一个累加器。

5.如权利要求3所述的装置,其中,每个所述处理字符串包括一个乘法器和一个累加器,以及

其中,所述处理阵列在所述多个子集中的每一个子集中包括逐元素操作处理器。

6.如权利要求1所述的装置,其中,将所述控制器进一步配置为:

遍历所述第一内存中的一个或多个批次的数据,以确定所述一个或多个批次的数据的尺寸覆盖对应于所述多个工作项中的每个工作项的尺寸的预定数据尺寸。

7.如权利要求6所述的装置,其中,将所述控制器进一步配置为:

当确定所述一个或多个批次的数据的尺寸不覆盖对应于所述多个工作项中的每个工作项的尺寸的预定数据尺寸时,将另一批次的数据提取到所述第一内存中。

8.如权利要求1所述的装置,其中,将所述控制器进一步配置为:

当确定所述一个或多个批次的数据中的部分数据在预定时间段内不使用时,释放所述一个或多个批次的数据中的部分数据。

9.如权利要求1所述的装置,其中,所述多个工作项中的每一个工作项具有第一数据尺寸,所述一个或多个批次的数据具有多个信道,并且每个信道具有覆盖所述第一数据尺寸的第二数据尺寸。

10.一种用于执行卷积神经网络操作的方法,包括:

将一个或多个批次的数据提取到第一内存中;

将所述一个或多个批次的数据重新分组为多个工作项,其中,第一工作项部分地重叠于所述多个工作项中的一个或多个工作项;以及

将所述多个工作项广播到处理阵列,其中,将所述第一工作项传送到所述处理阵列的两个或多个处理字符串。

11.如权利要求10所述的方法,其中,将所述多个处理字符串分类为多个子集,并且将所述第一工作项传送到所述多个子集中的每个子集的第一处理字符串。

12.如权利要求11所述的方法,还包括:

将多个滤波器传送到所述处理阵列，

其中，所述多个滤波器的数目对应于所述多个子集的数目，并且将所述多个滤波器中的每一个滤波器传送到所述多个子集中的对应子集。

13.如权利要求10所述的方法，还包括：

对所述两个或多个处理字符串中的第一工作项并行执行乘法运算。

14.如权利要求13所述的方法，还包括：

对所述两个或多个处理字符串中的乘法结果并行执行加法运算。

15.如权利要求10所述的方法，还包括：

遍历所述第一内存中的所述一个或多个批次的数据，以确定所述一个或多个批次的数据的尺寸覆盖对应于所述多个工作项中的每个工作项的尺寸的预定数据尺寸。

16.如权利要求15所述的方法，还包括：

当确定所述一个或多个批次的数据的尺寸不覆盖对应于所述多个工作项中的每个工作项的尺寸的预定数据尺寸时，将另一批次的数据提取到所述第一内存中。

17.如权利要求10所述的方法，还包括：

当确定所述一个或多个批次的数据中的部分数据在预定时间段内不使用时，释放所述一个或多个批次的数据中的部分数据。

18.如权利要求10所述的方法，还包括：

通过所述多个处理字符串并行地生成多个输出。

19.一种非暂时性计算机可读存储介质，存储可由计算装置的至少一个处理器执行的指令集，以使所述计算装置执行用于执行卷积神经网络操作的方法，所述方法包括：

将一个或多个批次的数据提取到第一内存中；

将所述一个或多个批次的数据重新分组为多个工作项，其中，第一工作项部分地重叠于所述多个工作项中的一个或多个工作项；以及

将所述多个工作项广播到处理阵列，其中，将所述第一工作项传送到所述处理阵列的两个或多个处理字符串。

20.如权利要求19所述的计算机可读存储介质，其中，将所述多个处理字符串分类为多个子集，并且将所述第一工作项传送到所述多个子集中的每个子集的第一处理字符串。

21.如权利要求20所述的计算机可读存储介质，其中，所述指令集可由计算装置的至少一个处理器执行，以使所述计算装置进一步执行：

将多个滤波器传送到所述处理阵列，

其中，所述多个滤波器的数目对应于所述多个子集的数目，并且将所述多个滤波器中的每一个滤波器传送到所述多个子集中的对应子集。

22.如权利要求19所述的计算机可读存储介质，其中，所述指令集可由计算装置的至少一个处理器执行，以使所述计算装置进一步执行：

对所述两个或多个处理字符串中的第一工作项并行执行乘法运算。

23.如权利要求22所述的计算机可读存储介质，其中，所述指令集可由计算装置的至少一个处理器执行，以使所述计算装置进一步执行：

对所述两个或多个处理字符串中的乘法结果并行执行加法运算。

24.如权利要求19所述的计算机可读存储介质，其中，所述指令集可由计算装置的至少

一个处理器执行,以使所述计算装置进一步执行:

遍历所述第一内存中的所述一个或多个批次的数据,以确定所述一个或多个批次的数据的尺寸覆盖对应于所述多个工作项中的每个工作项的尺寸的预定数据尺寸。

25.如权利要求24所述的计算机可读存储介质,其中,所述指令集可由计算装置的至少一个处理器执行,以使所述计算装置进一步执行:

当确定所述一个或多个批次的数据的尺寸不覆盖对应于所述多个工作项中的每个工作项的尺寸的预定数据尺寸时,将另一批次的数据提取到所述第一内存中。

26.如权利要求19所述的计算机可读存储介质,其中,所述指令集可由计算装置的至少一个处理器执行,以使所述计算装置进一步执行:

当确定所述一个或多个批次的数据中的部分数据在预定时间段内不使用时,释放所述一个或多个批次的数据中的部分数据。

27.如权利要求19所述的计算机可读存储介质,其中,所述指令集可由计算装置的至少一个处理器执行,以使所述计算装置进一步执行:

通过所述多个处理字符串并行地生成多个输出。

28.一种终端,包括:

宿主单元;以及

一种用于执行与所述宿主单元通信耦合的卷积神经网络操作的装置,所述装置包括:

第一内存;

包含多个处理字符串的处理阵列;以及

控制器,配置为:

将一个或多个批次的数据提取到所述第一内存中;

将一个或多个批次的数据重新分组为多个工作项,其中,第一工作项部分地重叠于所述多个工作项中的一个或多个工作项;以及

将所述多个工作项广播到所述处理阵列,其中,所述第一工作项被传送到所述处理阵列的两个或多个处理字符串。

卷积神经网络的数据重用与高效处理方法

背景技术

[0001] 机器学习已广泛应用于各个领域。卷积神经网络(CNN)是一种广泛应用于机器学习的神经网络。CNNs在图像处理、语音识别、游戏、机器人等领域有着广泛的应用,提高CNNs的处理效率对于提高神经网络的整体执行性能具有重要意义。

发明内容

[0002] 本发明的实施例提供了一种用于执行卷积神经网络操作的装置。所述装置包括第一内存、包含多个处理字符串的处理阵列和控制器。所述控制器可以被配置为将一个或多个批次的数据提取到所述第一内存中,将所述一个或多个批次的数据重新分组为多个工作项,其中,第一工作项部分地重叠于所述多个工作项中的一个或多个工作项,并且将所述多个工作项广播到所述处理阵列,其中,所述第一工作项被传送至所述处理阵列的两个或多个处理字符串。

[0003] 本发明的实施例还提供了一种用于执行卷积神经网络操作的方法。所述方法包括将一个或多个批次的数据提取到第一内存中,将所述一个或多个批次的数据重新分组为多个工作项,其中,第一工作项部分地重叠于所述多个工作项中的一个或多个工作项,并且将所述多个工作项广播到包含多个处理字符串的处理阵列,其中,所述第一工作项被传送至所述处理阵列的两个或多个处理字符串。

[0004] 本发明的实施例还提供了一种非暂时性计算机可读存储介质,所述存储介质存储有一组可由计算装置的至少一个处理器执行的指令,以使所述计算装置执行用于执行卷积神经网络操作的方法。所述方法包括将一个或多个批次的数据提取到第一内存中,将所述一个或多个批次的数据重新分组为多个工作项,其中,第一工作项部分地重叠于所述多个工作项中的一个或多个工作项,并且将所述多个工作项广播到包含多个处理字符串的处理阵列,其中,所述第一工作项被传送至所述处理阵列的两个或多个处理字符串。

[0005] 本发明的实施例还提供了一种终端,包括宿主单元和用于执行与所述宿主单元通信耦合的卷积神经网络操作的装置。所述装置包括第一内存、包含多个处理字符串的处理阵列以及控制器。所述控制器可以被配置为将一个或多个批次的数据提取到所述第一内存中,将所述一个或多个批次的数据重新分组为多个工作项,其中,第一工作项部分地重叠于所述多个工作项中的一个或多个工作项,并且将所述多个工作项广播到所述处理阵列,其中,所述第一工作项被传送至所述处理阵列的两个或多个处理字符串。

[0006] 所公开的实施例的附加特征和优点将在以下描述中部分地阐述,并且部分可以从所述描述中明显看出,或者可以通过实施例的实践而获知。所公开的实施例的特征和优点可以通过权利要求中阐述的要素和组合来实现和获得。

[0007] 需要理解的是,上述一般性描述和下面的详细描述都仅是示例性和说明性的,并且所公开的实施例不作为对所要求保护的内容的限定。

附图说明

[0008] 图1示出了示例性卷积神经网络(CNN)操作。

[0009] 图2A示出了与本发明的实施例相符的示例性神经网络加速器架构。

[0010] 图2B示出了与本发明的实施例相符的示例性神经网络加速器核心架构。

[0011] 图2C示出了与本发明的实施例相符的、包含神经网络加速器的示例性云系统的示意图,。

[0012] 图3示出了与本发明的实施例相符的示例性操作单元配置。

[0013] 图4示出了与本发明的实施例相符的示例性控制器的框图。

[0014] 图5示出了用于CNN操作的输入数据的示例。

[0015] 图6A示出了与本发明的实施例相符的在第一时间周期内提取数据的示例。

[0016] 图6B示出了与本发明的实施例相符的在第二时间周期内提取和组合数据的示例。

[0017] 图6C示出了与本发明的实施例相符的在第三时间周期内提取和组合数据的示例。

[0018] 图7示出了与本发明的实施例相符的用于执行卷积神经网络操作的示例性流程图。

具体实施方式

[0019] 现在将详细参考示例性实施例,所述实施例的示例在附图中示出。后续描述参考所述附图,除非另有说明,否则不同附图中的相同编号表示相同或相似的元件。在以下示例性实施例的描述中阐述的实施方式并不代表符合本发明的所有实施方式。相反,它们仅仅作为与本发明所附权利要求书中所列举的相关方面相符的装置和方法的示例。

[0020] 图1示出了示例性卷积神经网络(CNN)操作。在该示例性操作中,诸如激活之类的输入数据102被结构化为跨多个(例如,C)信道的一组二维(2D)特征图。每个所述二维特征图能够涉及一个信道。如图1所示,输入数据102(例如,一幅图像)具有C数量的特征图,并且输入数据102的一个信道的尺寸为H×W。因此,所述输入数据102的尺寸可以为H×W×C。

[0021] 在图1中,可以采用滤波器104对输入数据102进行卷积。输入数据102的不同特征图可以具有诸如权重、偏差项等不同参数,然而同一特征图可以共享相同的参数。因此,每个滤波器104可以具有与输入数据102的C数量的特征图相对应的C数量的信道。滤波器104的每个信道可在输入数据102的对应特征图上滑动。如图1所示,滤波器104的每个信道的尺寸为S×R,一个滤波器104的尺寸可以为S×R×C。在此,在用于卷积运算的输入数据102上滑动的窗口的尺寸可以为S×R。在该示例中,输入数据102与K数量的滤波器104_1至104_k进行卷积。

[0022] 当第一滤波器104_1的第一信道在输入数据102的第一特征图上滑动以进行卷积运算时,将第一滤波器104_1的第一信道乘以输入数据102的第一特征图中的诸如b1至b3的接收域。可以将接收域b1至b3定义为与相邻的接收域部分重叠。例如,如图1所示,第一接收域b1与第二接收域b2和第三接收域b3部分重叠。输入数据102的其余特征图的接收域可以对应于所述第一特征图的接收域进行定义。因此,每个第一接收域b1至第三接收域b3均具有C个信道数。出于说明的目的,在本发明中,当输入数据102的每个特征图具有B数量的接收域时,可以认为所述输入数据102包括B数量的工作项,每个工作项都包括C数量的信道。在此,每个工作项的C个信道可以具有与所述接收域的尺寸S×R相对应的尺寸。

[0023] 可以通过将第一滤波器104_1与输入数据102的第一接收域b1相乘、并将C数量的信道的相乘结果求和来生成一个输出值。例如,通过将第一滤波器104_1的每个信道与输入数据102的第一接收域b1中的对应特征图相乘,并且将来自C个信道的相乘结果求和,可以生成第一输出值r1_c1。通过逐个信道地将输入数据102的第一滤波器104_1和第二接收域b2相乘,并通过对C个信道的相乘结果求和,可以生成第二输出值r2_c1。通过将第一滤波器104_1在输入数据102上滑动而生成的包括第一输出值r1_c1和第二输出值r1_c2的B数量的输出值可以构成输出数据106的第一信道。

[0024] 类似地,可以通过对输入数据102卷积第二滤波器104_2来生成B数量的输出值,并且可以构成输出数据106的第二信道。也可以通过对输入数据102卷积第k个滤波器104_k来生成B数量的输出值,并且可以构成输出数据106的第k个信道。例如,可以通过将第K个滤波器104_k的每个信道与输入数据102的第一接收域b1中的对应特征图相乘,并通过将C数量的信道的乘法结果求和来生成第k个信道上的第一输出值r1_ck。如图1所示,输出数据106可以具有与滤波器104的数目相对应的K个信道,并且每个信道的尺寸为H′×W’。因此,输出数据106的尺寸可以是H’×W’×K。在该示例中,H’×W’可以等于输入数据102的工作项的数量,即B。在某些实施例中,输出数据106可以是卷积运算的中间输出数据。在某些实施例中,输出数据106可以通过包括逐元素操作的其他操作进一步处理,以生成用于所述卷积运算的最终输出数据。

[0025] 当执行卷积运算时,输入数据102的每个接收域(例如b1至b3)的数据从片上或片外内存提取到缓冲存储器中进行计算。如上所述,接收域b1至b3部分相互重叠。接收域b1至b3之间的重叠数据通常从片上内存或片外内存中提取,并多次存储到缓冲存储器中进行卷积运算,这会导致缓冲区空间不足或数据传送延迟。因此,对于接收域b1到b3之间的重叠数据的数据重用或共享方案可以通过减少存储在缓冲器中的数据或通过最小化数据传送带宽的使用来提高总体系统吞吐量。本发明的实施例可以提供一种能够有效地处理CNN操作的加速器。本发明的实施例还可以提供适于执行CNN操作的数据重用或共享方案。

[0026] 图2A示出了与本发明实施例相符的示例性神经网络加速器架构。在本发明的上下文中,神经网络加速器也可被称为机器学习加速器或深度学习加速器。在某些实施例中,加速器架构200可被称为神经网络处理单元(NPU)架构200。如图2A所示,加速器架构200可以包括多个核心202、命令处理器204、直接内存访问(DMA)单元208、联合测试操作组(JTAG)/测试访问端(TAP)控制器210、外围接口212、总线214等。

[0027] 可以理解的是,核心202可以基于所传送的数据执行算法操作。核心202可包括一个或多个处理元件,所述处理元件可包括单指令多数据(SIMD)架构,该架构包括一个或多个处理单元,所述处理单元被配置为基于从命令处理器204接收到的命令来执行一个或多个操作(例如,乘法、加法、乘法累加等)。为了对所通信的数据包执行操作,核心202可以包括用于处理数据包中的信息的一个或多个处理元件。每个处理元件可以包括任意数量的处理单元。根据本发明的某些实施例,加速器架构200可以包括多个核心202,例如,四个核心。在某些实施例中,多个核心202可以彼此通信耦合。例如,多个核心202可与单向环形总线连接,其支持用于大型神经网络模型的高效流水线。核心202的架构将参照图2B进行详细说明。

[0028] 命令处理器204可以与宿主单元220交互,并将相关命令和数据传递到相应的核心

202。在某些实施例中,命令处理器204可以在内核模式驱动器(KMD)的监督下与宿主单元进行交互。在某些实施例中,命令处理器204可以修改针对每个核心202的相关命令,使得核心202可以尽可能地并行工作。修改后的命令可以存储在指令缓冲器中。在某些实施例中,命令处理器204可以被配置为协调一个或多个核心202以用于并行执行。

[0029] DMA单元208可以辅助在主机存储器221与加速器架构200之间传送数据。例如,DMA单元208可以辅助将数据或指令从主机存储器221加载到核心202的本地内存中。DMA单元208还可以辅助在多个加速器之间传送数据。DMA单元208可以允许片外设备访问片内和片外内存而不引起主机CPU中断。另外,DMA单元208可以辅助在加速器架构200的组件之间传送数据。例如,DMA单元208可以辅助在多个核心202之间或每个核心内传送数据。因此,DMA单元208还可以生成内存地址并启动内存读取或写入周期。DMA单元208还可以包含可由一个或多个处理器写入和读取的数个硬件寄存器,包括内存地址寄存器、字节计数寄存器、一个或多个控制寄存器以及其他类型的寄存器。这些寄存器可以指定源、目的地、传送方向(从输入/输出(I/O)设备读取或写入I/O设备)、传送单元的尺寸或在一个突发中传送的字节数的某种组合。应当理解,加速器架构200可以包括第二DMA单元,该第二DMA单元可用于在其它加速器架构之间传送数据,以允许多个加速器架构在不涉及主机CPU的情况下直接通信。

[0030] JTAG/TAP控制器210可以指定一个专用的调试端口实现串行通信接口(例如,JTAG接口),用于对所述加速器的低开销访问,而无需直接从外部访问系统地址和数据总线。JTAG/TAP控制器210还可以具有片上测试访问接口(例如,TAP接口),所述接口实现访问一组测试寄存器的协议,所述测试寄存器呈现了各个部件的芯片逻辑位准和设备能力。

[0031] 外围接口212(例如PCIe接口)如果存在的话,(通常是)用作芯片间总线,用于在所述加速器与其他设备之间提供通信。

[0032] 总线214(例如I2C总线)包括芯片内总线和芯片间总线。芯片内总线按照系统架构的要求将所有内部组件彼此连接起来。虽然并非所有组件都连接到其他组件,但所有组件都与它们需要与之通信的其他组件建立了某种连接。芯片间总线将加速器与其他设备(例如片外存储器或外围设备)连接起来。例如,总线214可以跨核心提供高速通信,并且还可以将核心202与其它单元(例如片外内存或外围设备)连接。通常,如果存在外围接口212(例如,芯片间总线),则总线214仅与芯片内总线有关,尽管在某些实现中,总线214仍然可以涉及专用的总线间通信。

[0033] 加速器架构200也可以与宿主单元220通信。宿主单元220可以是一个或多个处理单元(例如,X86中央处理单元)。如图2A所示,宿主单元220可以与主机存储器221相关联。在某些实施例中,主机存储器221可以是与宿主单元220相关联的整体存储器或外部存储器。在某些实施例中,主机存储器221可以包括主机磁盘,主机磁盘是配置成为宿主单元220提供附加内存的外部存储器。主机存储器221可以是双倍数据速率同步动态随机存取内存(例如,DDR、SDRAM)或诸如此类。与集成在加速器芯片中的片上内存相比,主机存储器221可以被配置成以较慢的存取速度存储大量数据,以此作为更高级别的缓存。存储在主机存储器221中的所述数据可以被传送到加速器架构200以用于执行神经网络模型。

[0034] 在某些实施例中,具有宿主单元220和主机存储器221的主机系统可以包括编译器(未示出)。编译器是一种程序或计算机软件,它将用一种编程语言编写的计算机代码转换

为用于供加速器架构200创建可执行程序的指令。在机器学习应用中,编译器可以执行各种操作,例如,预处理、词法分析、语法分析、语义分析、将输入程序转换为中间表示、神经网络初始化、代码优化以及代码生成,或其组合。例如,编译器可以编译一个神经网络以生成静态参数,例如,神经元之间的连接和所述神经元的权重。

[0035] 在某些实施例中,包括编译器的主机系统可以将一个或多个命令推送到加速器架构200。如上所述,这些命令可以由加速器架构200的命令处理器204进一步处理,临时存储在加速器架构200的指令缓冲器中,并分发到相应的一个或多个核心(例如,图2A中的核心202)或处理元件中。某些命令可以指示DMA单元(例如,图2A的DMA单元208)将指令和数据从主机存储器(例如,图2A的主机存储器221)加载到加速器架构200中。然后,可以将加载的指令分发到分配有所述相应任务的每个核心(例如,图2A的核心202),并且所述一个或多个核心可以处理这些指令。

[0036] 应当理解的而是,由所述核心202接收的前几个指令可以指示所述核心202将来自主机存储器221的数据加载/存储到所述核心的一个或多个本地内存(例如,图2B的本地内存2032)。接着每个核心202可启动指令管道,所述指令管道包括从指令缓冲器提取指令(例如,通过序列器),解码所述指令(例如,通过图2A的DMA单元208),生成本地内存地址(例如,对应于操作数),读取源数据,执行或加载/存储操作,然后写回结果。

[0037] 根据某些实施例,加速器架构200可以进一步包括全局内存(未示出),该全局内存具有用作主内存的内存块(例如,8GB第二代高带宽内存(HBM2)的4个块)。在某些实施例中,全局内存可以经由DMA单元208存储来自主机存储器221的指令和数据。然后,这些指令可以被分发到分配有相应任务的每个核心的指令缓冲器,并且所述核心可以相应地处理这些指令。

[0038] 在某些实施例中,加速器架构200可进一步包括内存控制器(未示出),其被配置为对从全局内存内的特定内存块(例如HBM2)读取数据以及向该特定内存块写入数据进行管理。例如,内存控制器可以管理来自另一个加速器的核心(例如,来自DMA单元208或对应于另一个加速器的DMA单元)或来自核心202(例如,来自核心202中的本地内存)的数据读取/写入。应当理解,可以在加速器架构200中提供一个以上的内存控制器。例如,对于全局内存中的每个内存块(例如,HBM2),可以有一个内存控制器。

[0039] 内存控制器可以生成内存地址并初始化内存读取或写入周期。内存控制器可以包含数个硬件寄存器,这些寄存器可以由所述一个或多个处理器写入和读取。所述寄存器可以包括内存地址寄存器、字节计数寄存器、一个或多个控制寄存器以及其他类型的寄存器。这些寄存器可以指定源、目的地、传送方向(从输入/输出(I/O)设备读取或写入所述I/O设备)、所述传送单元的尺寸、在一个突发中传送的字节数、或内存控制器的其他典型功能的某种组合。

[0040] 虽然在本发明的某些实施例中,图2A的加速器架构200可用于卷积神经网络(CNNs),但是应理解的是,图2A的加速器架构200可以用于各种神经网络,例如深层神经网络(DNNs)、递归神经网络(RNNs)、或诸如此类。另外,某些实施例可以被配置为各种处理架构,例如神经网络处理单元(NPUs)、图形处理单元(GPUs)、现场可编程门阵列(FPGAs)、张量处理单元(TPUs)、专用集成电路(ASICs),任何其他类型的异构加速器处理单元(HAPUs)、或诸如此类。

[0041] 图2B示出了与本发明实施例相符的示例性核心架构。如图2B所示,核心202可以包括一个或多个操作单元,例如第一和第二操作单元2020和2022、内存引擎2024、序列器2026、指令缓冲器2028、常量缓冲器2030、本地内存2032等。

[0042] 一个或多个操作单元可以包括第一操作单元2020和第二操作单元2022。第一操作单元2020可以被配置为对接收到的数据(例如,矩阵)执行操作。在某些实施例中,第一操作单元2020可以包括一个或多个处理单元,其被配置为执行一个或多个运算(例如,乘法、加法、乘法累加、逐元素运算等)。在某些实施例中,第一操作单元2020被配置为执行加速卷积运算或执行矩阵乘法运算。以下将参照图3详细说明第一操作单元2020的示例。

[0043] 第二操作单元2022可以被配置为执行池化操作、插值操作、关注区域(ROI)操作以及诸如此类。在某些实施例中,第二操作单元2022可以包括插值单元、池化数据路径以及诸如此类。

[0044] 内存引擎2024可以被配置为在对应的核心202内或两个核心之间执行数据复制。DMA单元208可以协助在对应的核心内或两个核心之间复制数据。例如,DMA单元208可以支持内存引擎2024将数据从本地内存(例如,图2B的本地内存2032)复制到相应的操作单元。内存引擎2024还可以被配置为执行矩阵转置,以使所述矩阵适于在所述操作单元中使用。

[0045] 序列器2026可与指令缓冲器2028耦合,并配置成检索命令,并将所述命令分发给核心202的组件。例如,序列器2026可以将卷积命令或乘法命令分发到第一操作单元2020,将池化命令分发到第二操作单元2022,或者将数据复制命令分发到内存引擎2024。序列器2026还可以被配置为监视神经网络任务的执行,并且使神经网络任务的子任务并行化,以提高执行的效率。在某些实施例中,第一操作单元2020、第二操作单元2022以及内存引擎2024可以根据存储在指令缓冲器2028中的指令、在序列器2026的控制下并行运行。

[0046] 指令缓冲器2028可以被配置为存储属于所述相应核心202的指令。在某些实施例中,指令缓冲器2028与序列器2026耦合,并且向序列器2026提供指令。在某些实施例中,存储在指令缓冲器2028中的指令可以由命令处理器204传送或修改。

[0047] 常量缓冲器2030可以被配置为存储常量值。在某些实施例中,存储在常量缓冲器2030中的常量值可以被诸如第一操作单元2020或第二操作单元2022之类的操作单元用来进行批量归一化、量化、反量化或类似的操作。

[0048] 本地内存2032可以提供具有快速的读/写速度的存储空间。为了尽可能减少与全局内存的交互,可以大容量的使用本地内存2032的存储空间。利用所述大容量的存储空间,大多数数据访问可以在核心202内执行,减少了数据访问引起的延迟。在某些实施例中,为了最小化数据加载延迟和能量消耗,集成在芯片上的SRAM(静态随机存取存储器)可以用作本地内存2032。在某些实施例中,本地内存2032可以具有192MB或更高的容量。根据本发明的某些实施例,本地内存2032均匀地分布在芯片上以减轻密集布线和发热问题。

[0049] 图2C示出了与本发明的实施例相符的包含加速器架构200的示例性云系统的示意图。如图2C所示,云系统230可以提供具有人工智能(AI)能力的云服务,并且可以包括多个计算服务器(例如,232和234)。在某些实施例中,例如,计算服务器232可以合并图2A的神经网络加速器架构200。为了简单和清楚起见,在图2C中以简化的方式示出了神经网络加速器架构200。

[0050] 借助于神经网络加速器架构200,云系统230可以提供图像识别、面部识别、翻译、

三维建模等扩展的人工智能功能。应当理解的是,神经网络加速器架构200可以以其它形式被部署到计算装置。例如,神经网络加速器架构200还可以集成在计算装置中,例如智能电话、平板电脑和可穿戴设备。

[0051] 图3示出了与本发明的实施例相符的示例性操作单元配置。根据本发明的实施例,操作单元可以是第一操作单元(例如,图2中的第一操作单元2020)。操作单元2020可以包括第一缓冲器310、第二缓冲器320以及处理阵列330。

[0052] 第一缓冲器310可以被配置为存储输入数据(例如,图1中的输入数据102)。在某些实施例中,存储在第一缓冲器310中的数据可以为将用于在处理阵列330中执行的输入数据。在某些实施例中,可以从本地内存(例如,图2B中的本地内存2032)提取输入数据。第一缓冲器310可以被配置为支持将在处理阵列330中使用的数据的重用或共享。在某些实施例中,存储在第一缓冲器310中的输入数据可以是用于卷积操作的激活数据。将参照图6A到图6C详细说明用于第一缓冲器310的示例性数据重用或共享方案。

[0053] 第二缓冲器320可被配置为存储权重数据(例如,图1中的权重数据104)。在某些实施例中,存储在第二缓冲器320中的权重数据可用于处理阵列330中以供执行。在某些实施例中,可以从本地内存(例如,图2B中的本地内存2032)提取存储在第二缓冲器320中的权重数据。在某些实施例中,存储在第二缓冲器320中的权重数据可以是用于卷积操作的滤波器数据(例如,图1中的滤波器104)。

[0054] 根据本发明的某些实施例,存储在第二缓冲器320中的权重数据可以是压缩数据。例如,权重数据可以为修剪后的数据,以节省芯片上的内存空间。在某些实施例中,操作单元2020可进一步包括稀疏引擎390。稀疏引擎390可以被配置为对将在处理阵列330中使用的压缩权重数据进行解压缩。

[0055] 处理阵列330可以具有多个层(例如,对应于图1中的K数量的滤波器104)。根据本发明的实施例,处理阵列330的每一层可以包括多个处理字符串,这些处理字符串可以并行地执行计算。例如,包括在处理阵列330的第一层中的第一处理字符串可以包括第一乘法器340_1和第一累加器350_1,并且第二处理字符串可以包括第二乘法器340_2和第二累加器350_2。类似地,所述第一层中的第i个处理字符串可以包括第i个乘法器340_i和第i个累加器350_i。虽然处理阵列330所执行的计算将根据图1的操作作为示例进行说明,但值得注意的是,本发明将不限于图1所示的示例。

[0056] 在某些实施例中,乘法器340可以被配置为对分配的工作项执行乘法运算。例如,第一层中的第一乘法器340_1可以执行第一接收域b1与第一滤波器104_1之间的乘法运算,并且第一层中的第二乘法器340_2可以执行第二接收域b2与第一滤波器104_1之间的乘法运算。类似地,第一层中的第i个乘法器340_i可以在第i个接收域bi和第一滤波器104_1之间执行乘法运算。

[0057] 累加器350可以在同一处理字符串中对来自累加器350之前的乘法器340的乘法结果执行求和运算。例如,第一层中的第一累加器350_1可以对来自第一乘法器340_1的乘法结果执行求和运算,并产生第一输出值r1_c1。第一层中的第二累加器350_2可以对来自第二乘法器340_2的乘法结果执行求和运算,并产生第二输出值r2_c1。类似地,第一层中的第i个累加器350_i可以对来自第i个乘法器340_i的乘法结果执行求和运算,并产生第i个输出值ri_c1。

[0058] 根据本发明的实施例,可以以类似方式配置处理阵列330的其他层,以执行与处理阵列330的第一层相似的功能。处理阵列330的第二层也可以具有多个处理字符串,每个处理字符串包括乘法器340和累加器350。在某些实施例中,处理阵列330的第二层中的处理字符串可以针对接收域b1至bi和第二滤波器104_2执行乘法运算和求和运算。例如,第二层中的第i个处理字符串的第i个乘法器340_i可以被配置为在第i个接收域bi与第二滤波器104_2之间执行乘法运算。第二层中的第i个处理字符串的第i个累加器350_i可以配置为对第二层第i个处理字符串的第i个乘法器340_i的乘法结果执行求和运算,并产生输出结果值ri_c2。类似地,处理阵列330的第K层中的处理字符串可以对接收域b1至bi与第K个滤波器104_k执行乘法运算和求和运算。例如,第k层第i个处理字符串的第i个乘法器340_i可以配置为在第i个接收域bi与第k个滤波器104_k之间执行乘法运算。第k层中的第i个处理字符串的第i个累加器350_i可以配置为对第K层中的第i个处理字符串的第i个乘法器340_i的乘法结果执行求和运算,并产生输出结果值ri_ck。

[0059] 在某些实施例中,处理阵列330可以在SIMD控制下执行计算。例如,当执行卷积操作(例如,图1中所示)时,处理阵列330的每一层可以使用不同的数据执行相同的指令。在图1所示的示例中,处理阵列330的第一层可以从第一缓冲器310接收与接收域b1至bi相对应的输入数据,并且从第二缓冲器320接收与第一滤波器104_1相对应的权重数据,并执行乘法与求和计算。处理阵列330的第二层可以接收与接收域b1到bi相对应的输入数据和与第二滤波器104_2相对应的权重数据,并执行乘法与求和计算。类似地,处理阵列330的第K层可以接收与接收域b1到bi相对应的输入数据和与第K个滤波器104_k相对应的权重数据,并执行乘法与求和计算。在所述示例中,处理阵列330的每一层可以使用相同的激活数据(例如接收域b1至bi)和不同的权重数据(例如,第一滤波器104_1至第K滤波器104_k)来执行与乘法运算和求和运算相对应的相同指令。在某些实施例中,对处理字符串的数目k×i进行SIMD控制,并且K×i个输出值可以并行产生。

[0060] 根据本发明的某些实施例,图3所示的处理阵列330可以包括在核心中(例如,图2B中的核心202)。当包含在处理阵列330的一层中的处理字符串的数目(例如,处理字符串的数目i)小于工作项的数目(例如,图1中的工作项数目B)时,在某些实施例中,可以由处理阵列330执行i个工作项,并且随后可以由处理阵列330执行其余的工作项(B-i个工作项数目)。在某些其它实施例中,可以通过处理阵列330执行i个工作项,并且可以由另一个核心中的另一个处理阵列330执行其余的工作项。

[0061] 根据本发明的某些实施例,处理阵列330可进一步包括逐元素操作处理器360。在某些实施例中,逐元素操作处理器360可以定位在处理字符串的末尾。在某些实施例中,处理阵列330的每一层中的处理字符串可以共享逐元素操作处理器360。例如,在处理阵列330的第一层中的i数量的处理字符串可以共享逐元素操作处理器360。在某些实施例中,处理阵列330的第一层中的逐元素操作处理器360可以依次对累加器350_1至350_i(例如r1_c1到ri_c1)的每个输出值执行其逐元素操作。类似地,处理阵列330的第K层中的逐元素操作处理器360可以依次对累加器350_1至350_i(例如从r1_ck到ri_ck)的每个输出值执行其逐元素操作。在某些实施例中,逐元素操作处理器360可以被配置成执行多个逐元素操作。在某些实施例中,由逐元素操作处理器360执行的逐元素操作可包括诸如ReLU函数、Leaky ReLU函数、Sigmoid函数、Tanh函数或类似的激活函数。

[0062] 在某些实施例中,乘法器340或累加器350可以被配置为针对不同的数据类型执行其操作,而这些数据类型与逐元素操作处理器360执行其操作的数据类型不同。例如,乘法器340或累加器350可以被配置为对整数类型数据(例如Int 8、Int 16等诸如此类的数据)执行其操作,而逐元素操作处理器360可以对诸如FP24之类的浮点类型数据执行其操作。因此,根据本发明的某些实施例,处理阵列330可以进一步包括解量化器370和量化器380,其中逐元素操作处理器360位于两者之间。在某些实施例中,由于解量化器370和批处理归一化操作都可以通过具有常数的乘法运算和加法运算来执行,这些操作可以从常量缓冲器2030提供,因此,批处理归一化操作可以合并到解量化器370。在某些实施例中,批处理归一化操作和解量化操作可以通过编译器合并到一个操作中。如图3所示,常量缓冲器2030可以向解量化器370提供用于解量化或批量归一化的常数。

[0063] 图4示出了与本发明的实施例相符的示例性控制器的框图。参考图1和图3所讨论的,当处理CNN操作时,接收域(例如域b1)可以部分地与相邻的接收域(例如域b2至b3)重叠。根据本发明的实施例,控制器400可被配置为支持数据重用和共享方案,该方案可适用于执行CNN操作。在某些实施例中,控制器400可以是操作单元2020的一部分,或者可以与操作单元2020分离。在某些实施例中,控制器400可以是存储引擎2024的一部分。在某些实施例中,控制器400可以是第一缓冲器310的一部分,或者可以与第一缓冲器310分离。

[0064] 如图4所示,控制器400可以包括数据提取器410、汇编器420和广播器430。数据提取器410可以被配置为将数据提取到图3的第一缓冲器310中。在某些实施例中,可以将数据从本地内存2032提取到第一缓冲器310。汇编器420可以被配置为对数据提取器410所提取的数据进行重新分组,用于从所述数据形成多个工作项。例如,汇编器420可以重新分组存储在第一缓冲器310中的数据以形成多个接收域b1、b2等。广播器430可以被配置为将由所述汇编器420形成的工作项广播到图3所示的包含在处理阵列330中的相应处理字符串。以下通过参考图5、图6A、图6B和图6C进一步详细说明数据提取器410、汇编器420和广播器430的示例。

[0065] 图5示出了用于卷积操作的输入数据的示例。图5中所示的输入数据可以为图1中的用于卷积操作的输入数据102的一部分,并且图5仅出于简单性和说明性的目的示出了输入数据的第一信道。如图5所示,输入数据102可以包括多个激活值。在某些实施例中,每个激活值可以由输入数据102的像素表示。在某些实施例中,多个激活值可以表示为排列在矩阵中的多个像素。

[0066] 在图5中,输入数据被示为具有4行和8列的矩阵。在本发明中,将仅出于说明的目的解释以3×3的窗口尺寸和1像素的步长执行卷积操作的实施例。例如,第一接收域b1具有C数量的信道,并且每个信道覆盖由第一行和第三行以及第一列和第三列界定的9个像素1.1至3.3。在图5中,出于说明的目的,第一接收域b1覆盖的像素1.1到3.3被着色。在该示例中,通过沿行方向将3×3窗口从第一接收域b1移动1个像素来定义第二接收域b2。可以通过将所述输入数据102上的3×3窗口从相邻接收域沿行方向或列方向滑动1像素来定义其它接收域。例如,第二接收域b2具有C数量的信道,并且每个信道覆盖由第一行和第三行、第二列和第四列界定的9个像素。第三接收域b3可以覆盖由第二行和第四行以及第一列和第三列界定的9个像素。第四接收域b4可以覆盖由第二行和第四行以及第二列和第四列界定的9个像素。

[0067] 如图5所示,第一接收域b1和其它接收域(例如b2到b4)部分重叠。例如,第一接收域b1和第二接收域b2共享6个像素1.2、1.3、2.2、2.3、3.2以及3.3,第一接收域b1和第三接收域b3共享6个像素2.1、2.2、2.3、3.1、3.2以及3.3。另外,第一接收域b1和第四接收域b4共享4个像素2.2、2.3、3.2以及3.3。如果按照常规技术将每个接收域b1至b4提取到缓冲区中用于卷积操作,则重叠像素的数据将被重复提取,这将导致可用带宽减少、缓冲区空间不足以及执行延迟。

[0068] 现在参考图6A以说明在第一时间周期T1中提取并存储第一缓冲器310中的数据的示例。在某些实施例中,图4的数据提取器410可以被配置为从本地内存2032提取一批数据。在该示例中,可以在一个周期内提取与4×2像素尺寸相对应的第一批数据。例如,对应于像素1.1、1.2、2.1、2.2、3.1、3.2、4.1和4.2的第一批数据610可以从存储在图5所示的本地内存2032中的输入数据102提取。虽然数据提取器410可以根据可用带宽或系统要求、在一个或多个周期中提取任意数量或形状的数据,但是在本发明中将解释在一个周期中提取4×2尺寸的数据的实施例。

[0069] 在第一时间周期T1,存储在第一缓冲器310中的数据不覆盖3×3窗口尺寸,汇编器420不开始重新分组工作项。根据本发明的实施例,汇编器420可以遍历第一缓冲器310中提取和存储的数据,以确定存储在第一缓冲器310中的数据是否覆盖至少一个窗口尺寸,例如,在该示例中为3×3尺寸。当数据提取器410在一个周期内提取的数据的尺寸小于窗口尺寸(例如,3×3尺寸)时,汇编器420可以等待,直到存储在第一缓冲器310中的数据的尺寸等于或大于窗口的尺寸。在该示例中,汇编器420可以在从本地内存2032提取第一批数据610和第二批数据620(如图6B所示)之后开始组装工作项。

[0070] 图6B示出了与本发明的实施例相符的第二时间周期内的数据提取和组装的示例。在第二时间周期T2中,数据提取器410可以从存储在图5所示的本地内存2032中的输入数据102提取与像素1.3、1.4、2.3、2.4、3.3、4.3和4.4相对应的第二批数据620。由于存储在第一缓冲器310的包含第一批数据610和第二批数据620的数据能够覆盖窗口尺寸,所以汇编器420可以从存储在第一缓冲器310的数据开始形成多个工作项,例如,汇编器420可以重新分组存储在第一缓冲器310中的数据以形成四个接收域b1至b4。注意,在该示例中,可以从存储在第一缓冲器310中的4×4尺寸的数据中组合出4个3×3尺寸的接收域,而不需要重复提取共享数据。在图6B中,可以从存储在第一缓冲器310中的4×4尺寸的输入数据组装四个工作项(例如,如图5所示的接收域b1至b4)。

[0071] 根据本发明的某些实施例,广播器430可以将由汇编器420形成的工作项传送到相应的处理元件(例如,图3所示的处理阵列330)。在某些实施例中,广播器430可以将工作项传送到处理阵列330的每一层。例如,广播器430可以将第一接收域b1传送到处理阵列330的第一至第k个信道的第一乘法器340_1,并将第二接收域b2传送到处理阵列330的第一至第k个信道的第二乘法器340_2。类似地,广播器430可以将第3和第4接收域b3和b4传送到处理阵列330的第一至第k个信道的第3和第4乘法器340_3和340_4。

[0072] 图6C示出了与本发明的实施例相符的第三时间周期内的数据提取和组装的示例。在第三时间周期T3中,数据提取器410可以从存储在图5所示的本地内存2032中的输入数据102提取与1.5、1.6、2.5、2.6、3.6、4.5和4.6中的像素相对应的第三批数据630。汇编器420可以通过类似于第二时间周期T2中的遍历存储在第一缓冲器310中的数据的过程,从第二

批数据620和第三批数据630形成诸如接收域b5至b8的工作项。注意,在该示例中,可以从存储在第一缓冲器310中的4×4尺寸的数据组装4个3×3尺寸的接收域b5至b8。

[0073] 此处,由于第一批数据610不再用于形成工作项,第一批数据610可以被释放或者可以被确定为从第一缓冲器310被释放。根据本发明的某些实施例,从第一缓冲器310释放数据可以包括从第一缓冲器310删除数据。如果第一批数据610将在后续时间段中使用,则可以将第一批数据610保持在第一缓冲器310中。根据本发明的某些实施例,为了防止本地内存2032和第一缓冲器310之间的额外数据传送,可以将不再被汇编器420使用的数据在第一缓冲器310中保持预定的时间段,以防在不久的将来再次使用。还应注意,在第三时间周期T3中,第二批数据620被重用以形成新的工作项,而无需再次提取第二批数据620。

[0074] 在第三时间周期T3中,广播器430还可以将由汇编器420新形成的工作项传送到相应的处理元件(例如,图3所示的处理阵列330)。例如,广播器430可以将第五接收域b5传送到处理阵列330的第一到第k个信道中的每个信道的第五乘法器340_5,并且将第六接收域b6传送到处理阵列330的第一到第k个频道中的每个信道的第六乘法器340_6。类似地,广播器430可以将第7和第8接收域b7和b8传送到处理阵列330的第一到第k个信道中的每个信道的第7和第8乘法器340_7和340_8。

[0075] 根据本发明的实施例,在随后的时间段中,可以提取一批尺寸为4×2的数据,并且可以形成4个尺寸为3×3的工作项。如上所述,根据本发明的实施例,利用从本地内存2032提取到第一缓冲器310的相对少量的数据,可以组装相对大量的工作项。因此,可以节省第一缓冲器310上的资源,并且在某些实施例中,可以减小第一缓冲器310的尺寸。根据本发明的某些实施例,由于数据重用和共享方案,因此可以以比常规技术相对较小的带宽满足数据提取需求。

[0076] 根据本发明的实施例,可以获得用于神经网络推理的高效工作项组装和处理技术。本发明的实施例可以提供用于CNN操作的工作项之间的数据重用和共享方案。本发明的实施例可以提供一种加速器架构,该架构能够基于工作项之间的数据重复特性有效地处理CNN操作。本发明的实施例能使得用于数据提取的带宽使用减少,并提升系统吞吐量。本发明的实施例还能够实现关于带宽和缓冲空间的有效资源使用。本发明的实施例还防止本地内存和缓冲存储器之间的重复数据传送。

[0077] 图7示出了与本发明的实施例相符的用于执行卷积神经网络操作的示例性流程图。为便于说明,将参考图4、图5、图6A、图6B和图6C来描述用于执行图1所示的卷积神经网络操作的方法,为便于说明,将解释以3×3的窗口尺寸和1像素的步长执行卷积操作的实施例。

[0078] 在步骤S710,可以将数据提取到缓冲存储器中。步骤S710可以由例如数据提取器410等执行。在某些实施例中,可以将数据从本地内存2032提取到第一缓冲器310。图5示出了要提取到第一缓冲器310的数据的示例。在某些实施例中,可以从本地内存2032将一批数据提取到第一缓冲器310。在该示例中,可以在一个周期内提取与4×2像素尺寸相对应的第一批数据,如图6A中的第一时间周期T1所示。虽然可以根据可用带宽或系统要求在一个或多个周期内提取任意数量或形状的数据,但在本发明中将解释在一个周期内提取4×2数据尺寸的实施例。

[0079] 如图6A中第一时间周期T1所示,当存储在第一缓冲器310中的数据没有覆盖3×3

的窗口尺寸时,可以不开始数据的重新分组。根据本发明的实施例,可以遍历在第一缓冲器310中提取和存储的数据,以确定存储在第一缓冲器310中的数据是否覆盖至少一个窗口尺寸,例如在本示例中为3×3尺寸。在某些实施例中,直到存储在第一缓冲器310中的数据的尺寸等于或大于窗口的尺寸,才可以执行数据的重新分组。

[0080] 如图6B的时间周期T2所示,当在一个周期内提取到缓冲存储器的数据的尺寸小于窗口尺寸时(例如,3×3尺寸),可以从本地内存2032向第一缓冲器310提取第二批数据620。由于存储在第一缓冲器310中的包含第一批数据610和第二批数据620的数据覆盖了一个窗口尺寸,因此该方法可以继续执行步骤S720。在步骤S720,可以执行对提取的数据的重新分组,以从所提取的数据中形成多个工作项。步骤S720可由诸如汇编器420等执行。例如,在步骤S720,存储在第一缓冲器310中的数据可以被重新分组,以形成四个接收域b1至b4。值得注意的是,在该示例中,可以从存储在第一缓冲器310中的4×4尺寸的数据中组合出4个3×3尺寸的接收域,而无需重复提取共享数据。在图6B中,可以将存储在第一缓冲器310中的4×4尺寸的输入数据分组成四个工作项,例如图5所示的接收域b1至b4。

[0081] 在步骤S730,可以将在步骤S720形成的工作项广播到相应的处理元件(例如,图3所示的处理阵列330)。步骤S730可由诸如广播器430等执行。在某些实施例中,可以将在步骤S720形成的工作项传送到处理阵列330的每一层。例如,第一接收域b1可以被传送到处理阵列330的第一至第k个信道的第一乘法器340_1,并且第二接收域b2可以被传送到处理阵列330的第一至第k个信道的第二乘法器340_2。类似地,第三和第四接收域b3和b4可以被传送到处理阵列330的第一至第k个信道的第三和第四乘法器340_3和340_4。

[0082] 如图6C所示,在执行卷积操作时可以重复步骤S710、S720和S730。例如,如图6C中的第三时间周期T3所示,可以从存储在本地内存2032中的输入数据102中提取第三批数据630。类似于第二时间周期T2中的处理,在遍历存储在第一缓冲器310中的数据之后,可以形成诸如来自第二批数据620和第三批数据630的接收域b5至b8之类的工作项。值得注意的是,在该示例中,可以从存储在第一缓冲器310中的4×4尺寸的数据中组装4个3×3尺寸的接收域b5至b8。

[0083] 此处,由于第一批数据610不再用于形成工作项,所以所述方法可以进一步包括用于从第一缓冲器310释放或确定欲释放第一批数据610的步骤。如果第一批数据610将在后续时间段中使用,则可以将第一批数据610保持在第一缓冲器310中。根据本发明的某些实施例,为了防止本地内存2032和第一缓冲器310之间的额外数据传送,可以将不再用于重组的数据在第一缓冲器310中保持预定的时间段,以备近期重用。还需注意的是,在第三时间周期T3中,第二批数据620被重用以形成新的工作项,而无需再次提取第二批数据620。

[0084] 在第三时间周期T3中,可以将新形成的工作项传送到相应的处理单元(例如,图3所示的处理阵列330)。例如,第五接收域b5可以被传送到处理阵列330的第一至第k个信道中每个信道的第五乘法器340_5,并且第六接收域b6可以被传送到处理阵列330的第一至第k个信道中每个信道的第六乘法器340_6。类似地,第7和第8接收域b7和b8可以被传送到处理阵列330的第一至第k个信道中每个信道的第7和第8乘法器340_7和340_8。

[0085] 在随后的时间段中,还可以在执行卷积操作期间重复步骤S710、S720以及S730。例如,根据本发明的实施例,可以在每个时间周期内可以提取一批尺寸为4×2的数据,并且可以形成4个尺寸为3×3的工作项。新形成的工作项也可以传送到相应的处理字符串中。

[0086] 可以使用以下条款进一步描述所述实施例：

[0087] 1.一种用于执行卷积神经网络操作的装置,包括：

[0088] 第一内存；

[0089] 包含多个处理字符串的处理阵列;以及

[0090] 控制器配置为：

[0091] 将一个或多个批次的数据提取到所述第一内存中；

[0092] 将所述一个或多个批次的数据重新分组为多个工作项,其中,第一工作项部分地重叠于所述多个工作项中的一个或多个工作项;以及

[0093] 将所述多个工作项广播到所述处理阵列,其中,将所述第一工作项传送到所述处理阵列的两个或多个处理字符串。

[0094] 2.如条款1所述的装置,其中,将所述多个处理字符串分类为多个子集,并且将所述第一工作项传送到所述多个子集中的每个子集的第一处理字符串。

[0095] 3.如条款2所述的装置,进一步包括第二存储器,所述第二存储器存储多个滤波器,所述多个滤波器的数目对应于所述子集的数目。

[0096] 4.如条款1至条款3中任一条款所述的装置,其中,每个所述处理字符串包括一个乘法器和一个累加器。

[0097] 5.如条款3所述的装置,其中,每个所述处理字符串包括一个乘法器和一个累加器,以及

[0098] 其中,所述处理阵列在所述多个子集中的每一个子集中包括逐元素操作处理器。

[0099] 6.如条款1至条款5中任一条款所述的装置,其中,将所述控制器进一步配置为：

[0100] 遍历所述第一内存中的一个或多个批次的数据,以确定所述一个或多个批次的数据的尺寸覆盖对应于所述多个工作项中的每个工作项的尺寸的预定数据尺寸。

[0101] 7.如条款6所述的装置,其中,将所述控制器进一步配置为：

[0102] 当确定所述一个或多个批次的数据的尺寸不覆盖对应于所述多个工作项中的每个工作项的尺寸的预定数据尺寸时,将另一批次的数据提取到所述第一内存中。

[0103] 8.如条款1至条款7中任一条款所述的装置,其中,将所述控制器进一步配置为：

[0104] 当确定所述一个或多个批次的数据中的部分数据在预定时间段内不使用时,释放所述一个或多个批次的数据中的部分数据。

[0105] 9.如条款1至条款5中任一条款所述的装置,其中,所述多个工作项中的每一个工作项具有第一数据尺寸,所述一个或多个批次的数据具有多个信道,并且每个信道具有覆盖所述第一数据尺寸的第二数据尺寸。

[0106] 10.一种用于执行卷积神经网络操作的方法,包括：

[0107] 将一个或多个批次的数据提取到第一内存中；

[0108] 将所述一个或多个批次的数据重新分组为多个工作项,其中,第一工作项部分地重叠于所述多个工作项中的一个或多个工作项;以及

[0109] 将所述多个工作项广播到所述处理阵列,其中,将所述第一工作项传送到所述处理阵列的两个或多个处理字符串。

[0110] 11.如条款10所述的方法,其中,将所述多个处理字符串分类为多个子集,并且将所述第一工作项传送到所述多个子集中的每个子集的第一处理字符串。

[0111] 12.如条款11所述的方法,还包括:

[0112] 将多个滤波器传送到所述处理阵列,

[0113] 其中,所述多个滤波器的数目对应于所述多个子集的数目,并且将所述多个滤波器中的每一个滤波器传送到所述多个子集中的对应子集。

[0114] 13.如条款10至条款12中任一条款所述的方法,还包括:

[0115] 对所述两个或多个处理字符串中的第一工作项并行执行乘法运算。

[0116] 14.如条款13所述的方法,还包括:

[0117] 对所述两个或多个处理字符串中的乘法结果并行执行加法运算。

[0118] 15.如条款10至条款14中任一条款所述的方法,还包括:

[0119] 遍历所述第一内存中的所述一个或多个批次的数据,以确定所述一个或多个批次的数据的尺寸覆盖对应于所述多个工作项中的每个工作项的尺寸的预定数据尺寸。

[0120] 16.如条款15所述的方法,还包括:

[0121] 当确定所述一个或多个批次的数据的尺寸不覆盖对应于所述多个工作项中的每个工作项的尺寸的预定数据尺寸时,将另一批次的数据提取到所述第一内存中。

[0122] 17.如条款10至条款16中任一条款所述的方法,还包括:

[0123] 当确定所述一个或多个批次的数据中的部分数据在预定时间段内不使用时,释放所述一个或多个批次的数据中的部分数据。

[0124] 18.如条款10至条款17中任一条款所的方法,还包括:

[0125] 通过所述多个处理字符串并行地生成多个输出。

[0126] 19.一种非暂时性计算机可读存储介质,存储可由计算装置的至少一个处理器执行的指令集,以使所述计算装置执行用于执行卷积神经网络操作的方法,所述方法包括:

[0127] 将一个或多个批次的数据提取到第一内存中;

[0128] 将所述一个或多个批次的数据重新分组为多个工作项,其中,第一工作项部分地重叠于所述多个工作项中的一个或多个工作项;以及

[0129] 将所述多个工作项广播到所述处理阵列,其中,将所述第一工作项传送到所述处理阵列的两个或多个处理字符串。

[0130] 20.如条款19所述的计算机可读存储介质,其中,将所述多个处理字符串分类为多个子集,并且将所述第一工作项传送到所述多个子集中的每个子集的第一处理字符串。

[0131] 21.如条款20所述的计算机可读存储介质,其中,所述指令集可由计算装置的至少一个处理器执行,以使所述计算装置进一步执行:

[0132] 将多个滤波器传送到所述处理阵列,

[0133] 其中,所述多个滤波器的数目对应于所述多个子集的数目,并且将所述多个滤波器中的每一个滤波器传送到所述多个子集中的对应子集。

[0134] 22.如条款19至条款21中任一条款所述的计算机可读存储介质,其中,所述指令集可由计算装置的至少一个处理器执行,以使所述计算装置进一步执行:

[0135] 对所述两个或多个处理字符串中的第一工作项并行执行乘法运算。

[0136] 23.如条款22所述的计算机可读存储介质,其中,所述指令集可由计算装置的至少一个处理器执行,以使所述计算装置进一步执行:

[0137] 对所述两个或多个处理字符串中的乘法结果并行执行加法运算。

[0138] 24.如条款19至条款23中任一条款所述的计算机可读存储介质,其中,所述指令集可由计算装置的至少一个处理器执行,以使所述计算装置进一步执行:

[0139] 遍历所述第一内存中的所述一个或多个批次的数据,以确定所述一个或多个批次的数据的尺寸覆盖对应于所述多个工作项中的每个工作项的尺寸的预定数据尺寸。

[0140] 25.如条款24所述的计算机可读存储介质,其中,所述指令集可由计算装置的至少一个处理器执行,以使所述计算装置进一步执行:

[0141] 当确定所述一个或多个批次的数据的尺寸不覆盖对应于所述多个工作项中的每个工作项的尺寸的预定数据尺寸时,将另一批次的数据提取到所述第一内存中。

[0142] 26.如条款19至条款25中任一条款所述的计算机可读存储介质,其中,所述指令集可由计算装置的至少一个处理器执行,以使所述计算装置进一步执行:

[0143] 当确定所述一个或多个批次的数据中的部分数据在预定时间段内不使用时,释放所述一个或多个批次的数据中的部分数据。

[0144] 27.如条款19至条款26中任一条款所述的计算机可读存储介质,其中,所述指令集可由计算装置的至少一个处理器执行,以使所述计算装置进一步执行:

[0145] 通过所述多个处理字符串并行地生成多个输出。

[0146] 28.一种终端,包括:

[0147] 宿主单元;以及

[0148] 一种用于执行与所述宿主单元通信耦合的卷积神经网络操作的装置,所述装置包括:

[0149] 第一内存;

[0150] 包含多个处理字符串的处理阵列;以及

[0151] 控制器,配置为:

[0152] 将一个或多个批次的数据提取到所述第一内存中;

[0153] 将一个或多个批次的数据重新分组为多个工作项,其中,第一工作项部分地重叠于所述多个工作项中的一个或多个工作项;以及

[0154] 将所述多个工作项广播到所述处理阵列,其中,所述第一工作项被传送到所述处理阵列的两个或多个处理字符串。

[0155] 本文的实施例包括数据库系统、方法和有形的非暂时性计算机可读介质。所述方法可以被执行,例如由至少一个从有形的非暂时性计算机可读存储介质(诸如具有图2A的宿主单元220和主机存储器221的主机系统)接收指令的处理器执行。类似地,与本发明一致的系统可以包括至少一个处理器和存储器,并且所述存储器可以是有形的非暂时性计算机可读存储介质。如本文所使用的,有形的非暂时性计算机可读存储介质是指可在其上存储可由至少一个处理器读取的信息或数据的任何类型的物理存储器。示例包括随机存取存储器(RAM)、只读存储器(ROM)、易失性存储器、非易失性存储器、硬盘驱动器、CD-ROM、DVD、闪存驱动器、磁盘、寄存器、高速缓存和任何其他已知的物理存储介质。诸如“存储器”和“计算机可读存储介质”之类的单数术语可以附加地指代多种结构,如多个存储器或计算机可读存储介质。如本文所述,除非另有说明,“存储器”可包括任何类型的计算机可读存储介质。计算机可读存储介质可存储供至少一个处理器执行的指令,包括用于使处理器执行与本文实施例一致的步骤或阶段的指令。另外,一个或多个计算机可读存储介质可用于实现计算

机实现的方法。术语“非暂时性计算机可读存储介质”应理解为包括有形物品,并排除载波和瞬态信号。

[0156] 如本文所使用的,除非另有特别说明,术语“或”涵盖所有可能的组合,除非不可行。例如,如果声明数据库可以包括A或B,则除非另有特别说明或不可行,否则所述数据库可以包括A、或B、或A和B。作为第二个示例,如果声明数据库可以包括A、B、或C,则除非另有特别说明或不可行,否则所述数据库可以包括A、或B、或C、或者A和B、或者A和C,或者B和C,或者A和B和C。

[0157] 在上述详述中,已经参照可以随实施方式而变化的许多具体细节描述了实施例。能够对所述描述的实施例进行某些调整和修改。通过考虑本文公开的发明的说明书和实践,其他实施例对于本领域技术人员来说是显而易见的。所述说明书和实施例仅被视为示例性的,本发明的真正范围和精神由所附权利要求书指示。另外,图中所示的步骤顺序仅出于说明目的,并不旨在限定任何特定的步骤序列。因此,本领域技术人员可以理解,在实施相同方法的同时,可以以不同的顺序执行这些步骤。

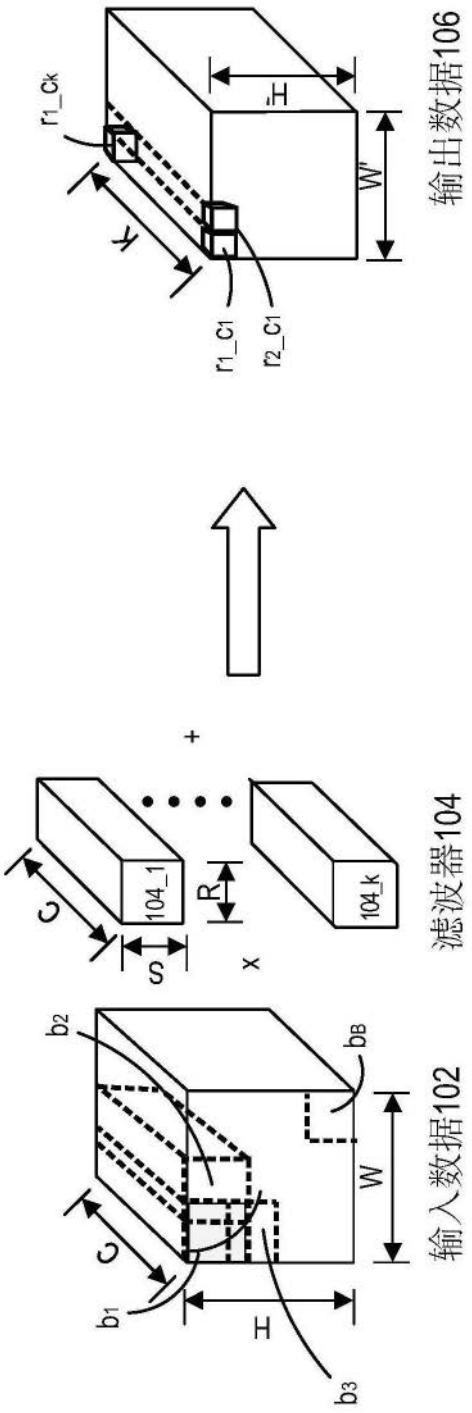

输出数据106
r1_ck
r1_c1
r2_c1
H'
W'
K
+
滤波器104
104_1
104_k
R
C
S
x
输入数据102
b2
bB
b1
b3
W
H
C

图1

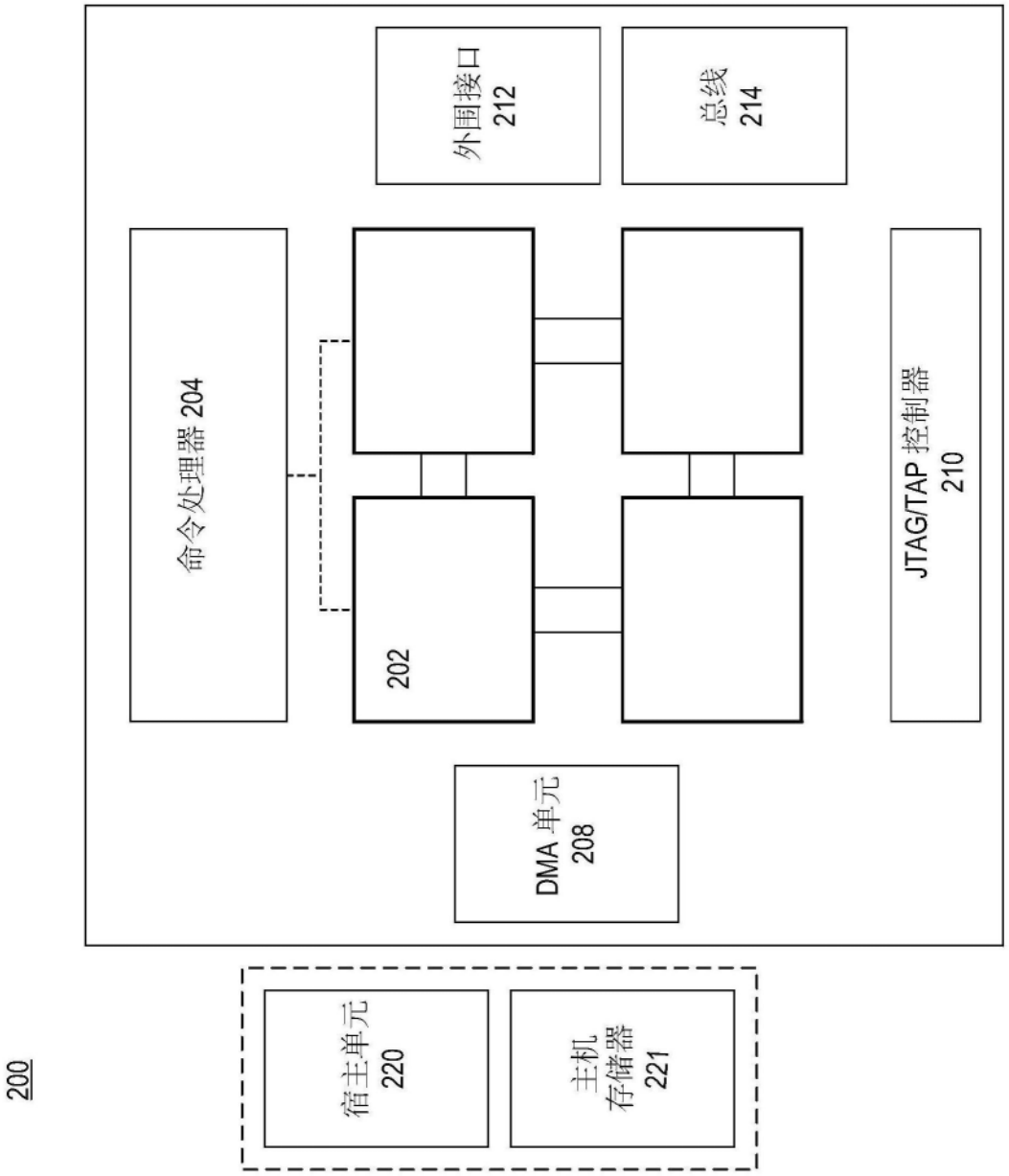

外围接口
212
总线
214
命令处理器 204
202
JTAG/TAP 控制器
210
DMA 单元
208
宿主单元
220
主机
存储器
221
200

图2A

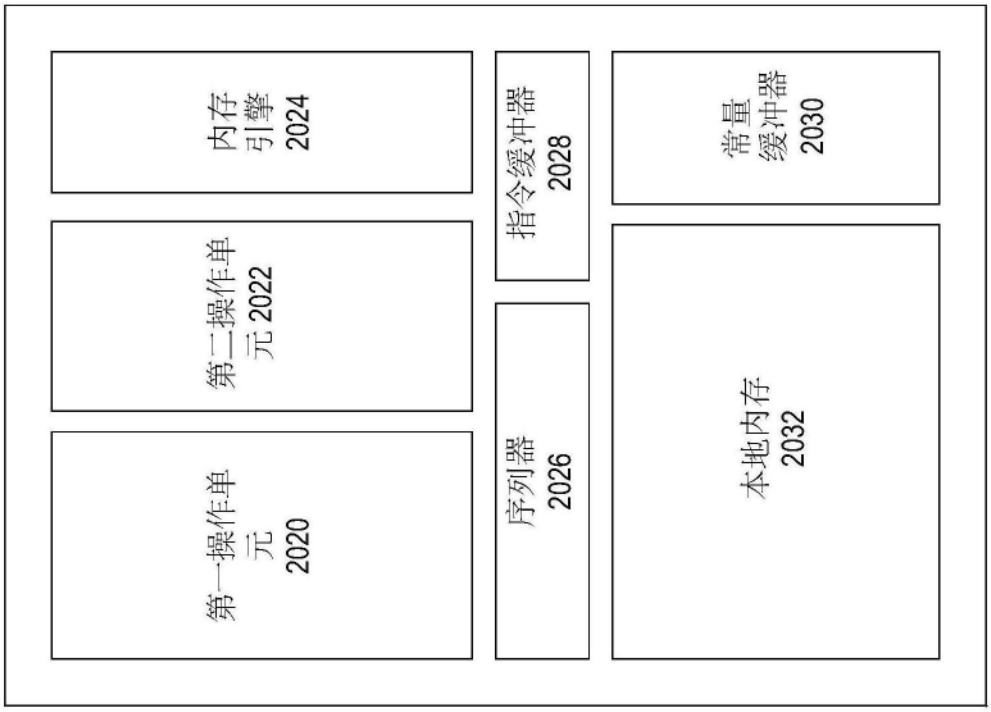

第一操作单元 2020
第二操作单元 2022
内存引擎 2024
序列器 2026
指令缓冲器 2028
本地内存 2032
常量缓冲器 2030
202

图2B

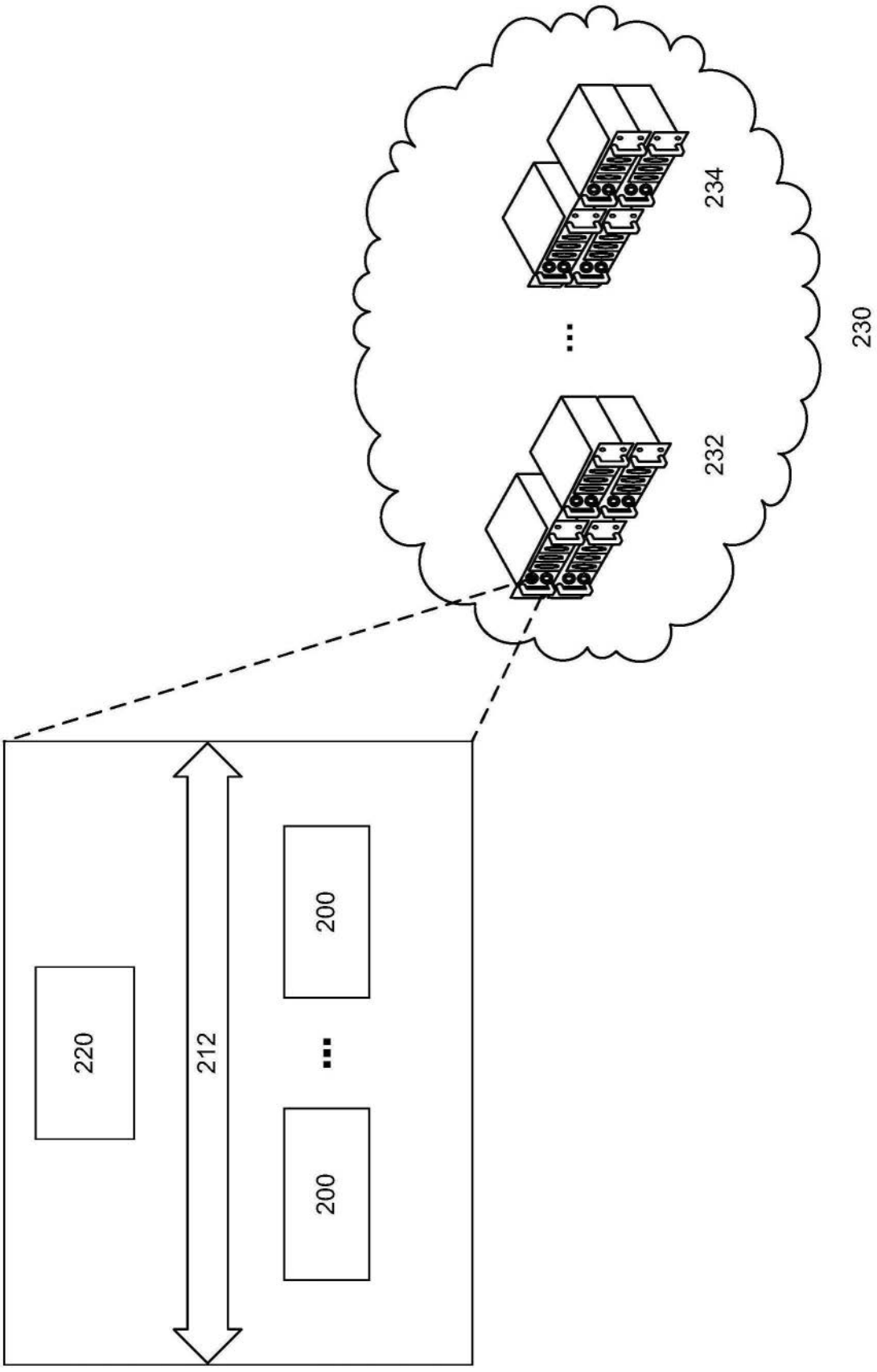


图2C

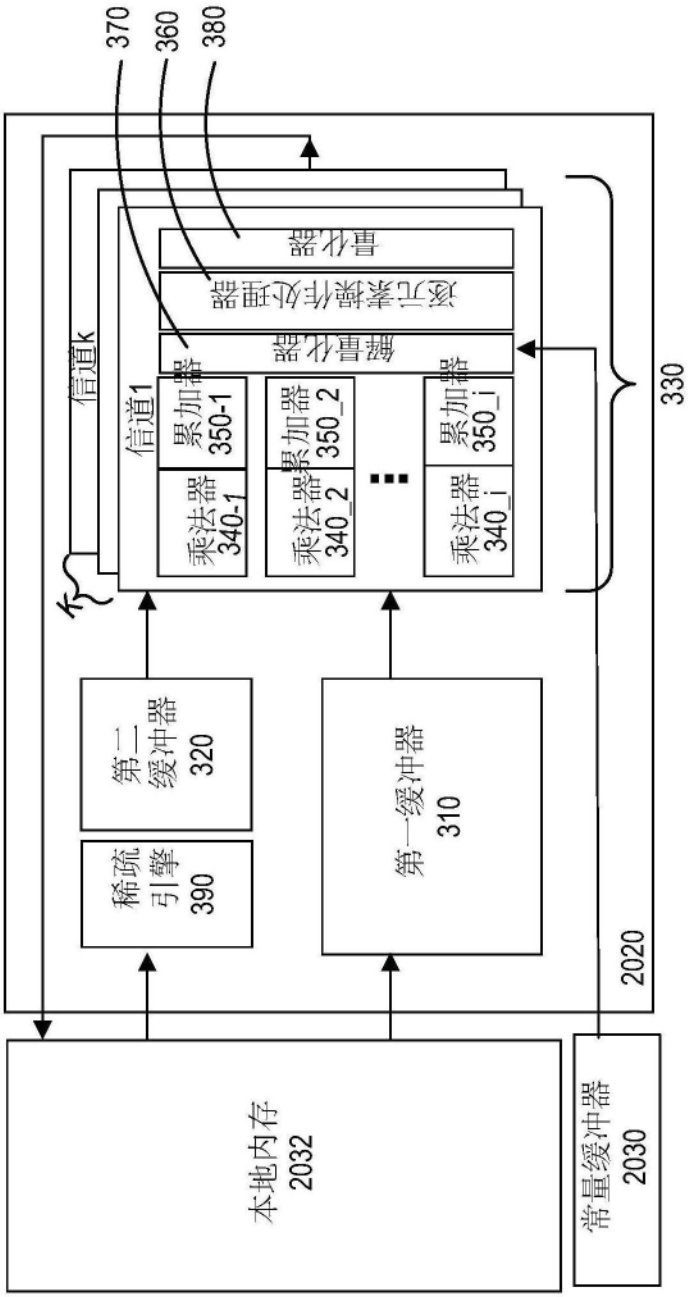

370
360
380
信道k
信道1
累加器 350-1
乘法器 340-1
累加器 350_2
乘法器 340_2
累加器 350_i
乘法器 340_i
330
第二缓冲器 320
稀疏引擎 390
第一缓冲器 310
2020
本地内存 2032
常量缓冲器 2030

图3

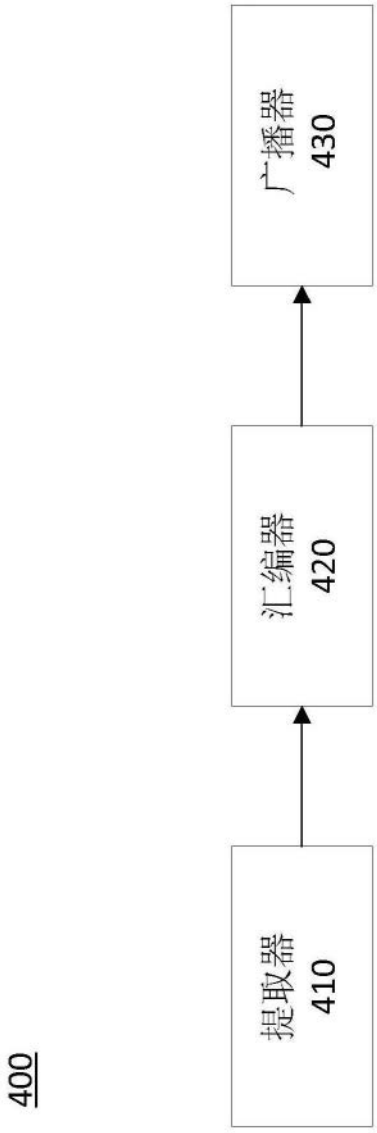


图4

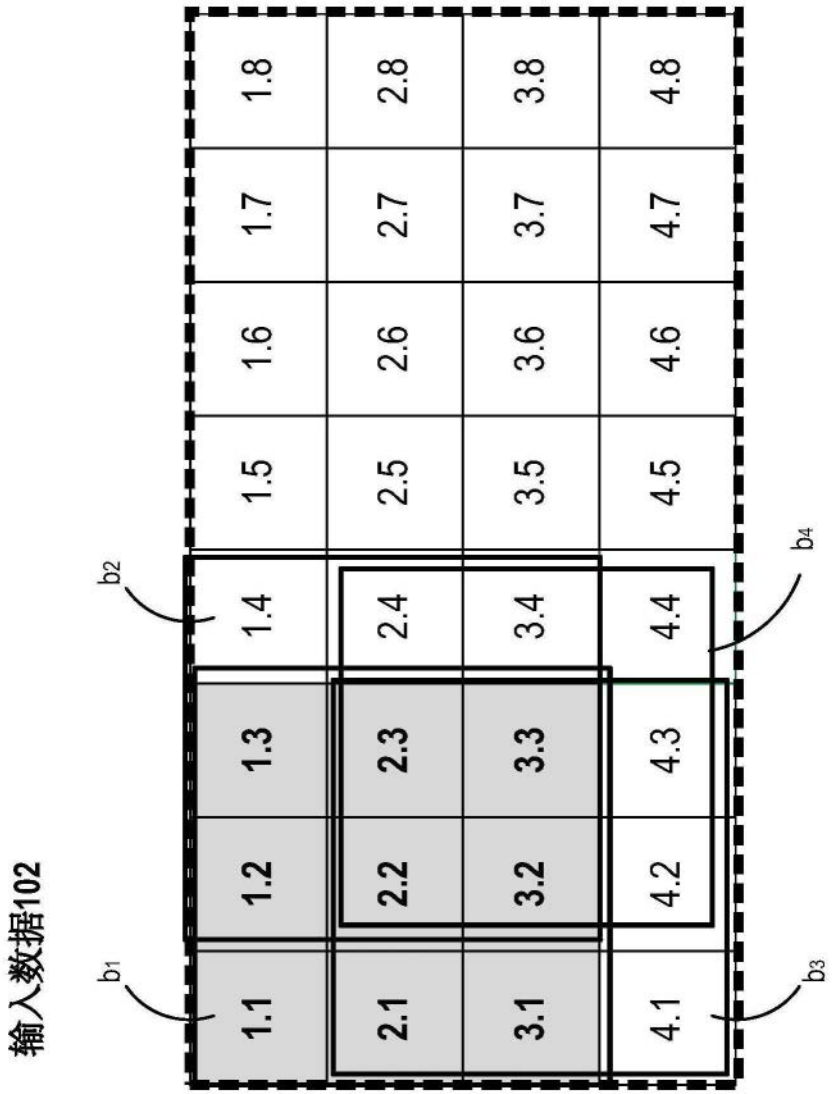

输入数据102
b1
b2
b3
b4
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
2.1
2.2
2.3
2.4
2.5
2.6
2.7
2.8
3.1
3.2
3.3
3.4
3.5
3.6
3.7
3.8
4.1
4.2
4.3
4.4
4.5
4.6
4.7
4.8

图5

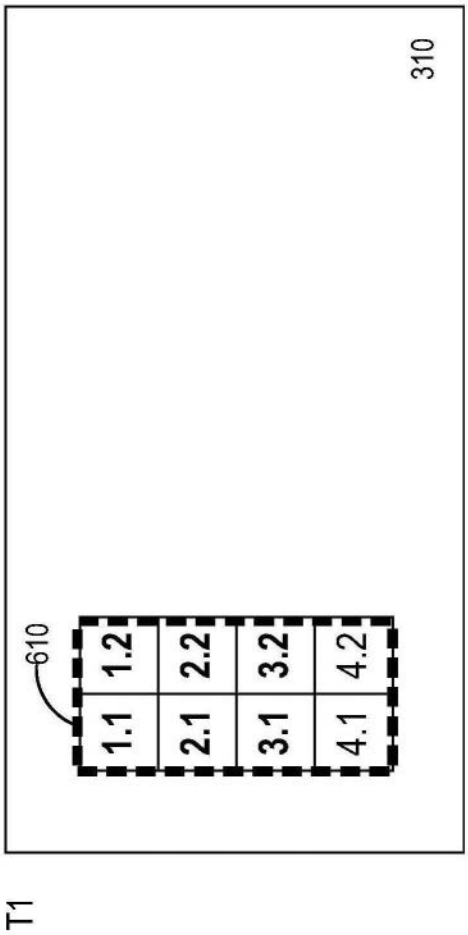


图6A

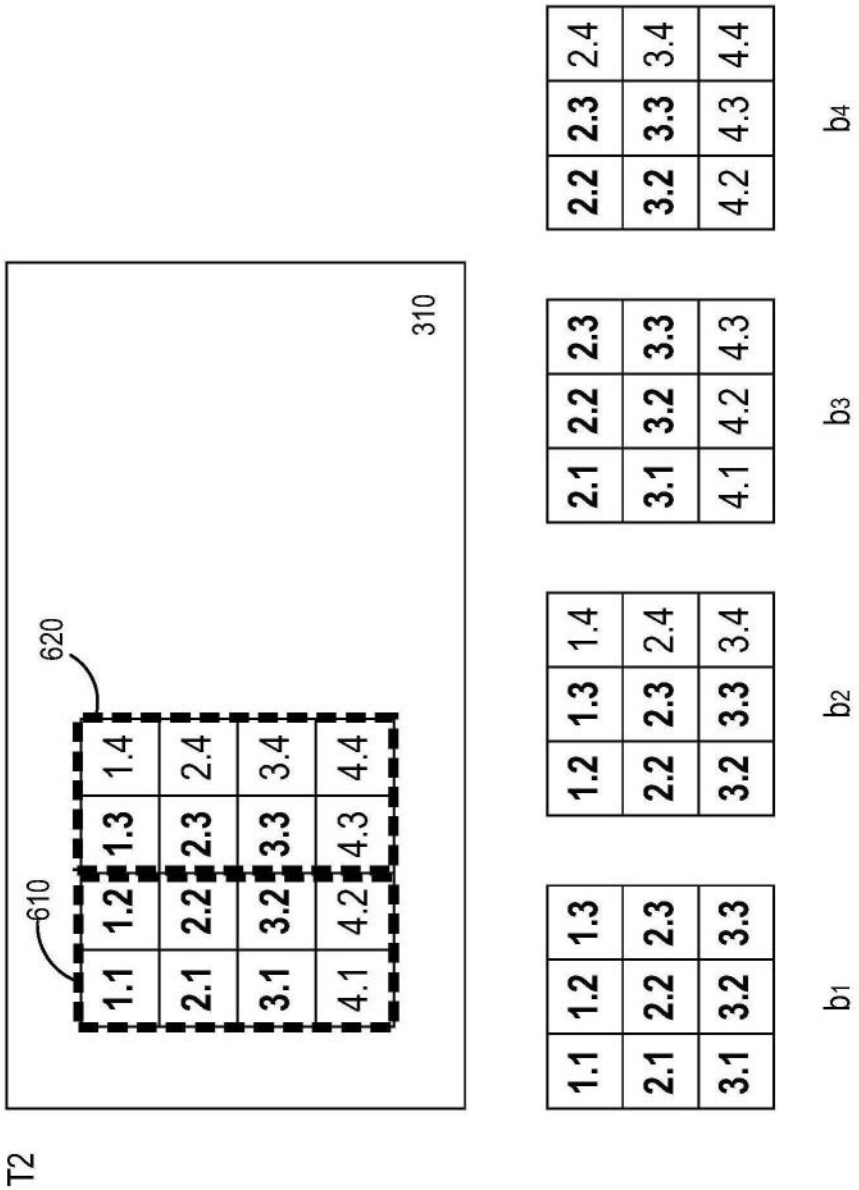

T2
310
620
610
1.1 1.2 1.3 1.4
2.1 2.2 2.3 2.4
3.1 3.2 3.3 3.4
4.1 4.2 4.3 4.4
b1
b2
b3
b4

图6B

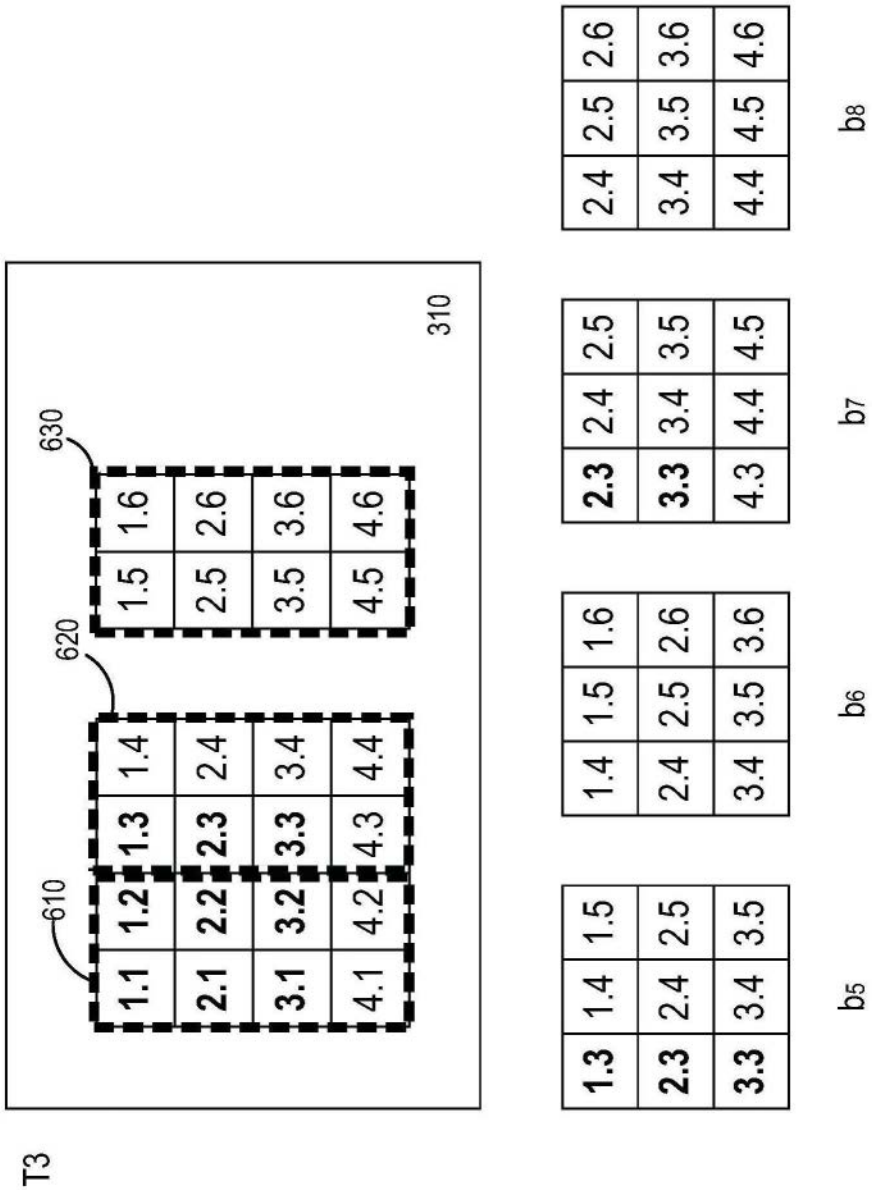

T3
310
630
620
610
1.1 1.2 1.3 1.4
2.1 2.2 2.3 2.4
3.1 3.2 3.3 3.4
4.1 4.2 4.3 4.4
1.5 1.6
2.5 2.6
3.5 3.6
4.5 4.6
1.3 1.4 1.5
2.3 2.4 2.5
3.3 3.4 3.5
b5
1.4 1.5 1.6
2.4 2.5 2.6
3.4 3.5 3.6
b6
2.3 2.4 2.5
3.3 3.4 3.5
4.3 4.4 4.5
b7
2.4 2.5 2.6
3.4 3.5 3.6
4.4 4.5 4.6
b8

图6C

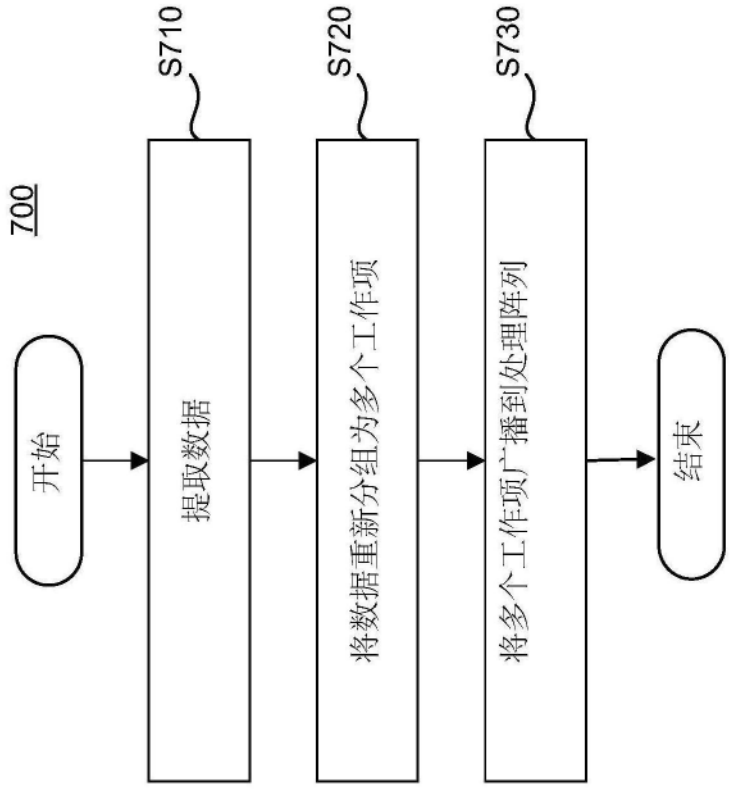


图7