(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 115936064 B
(45)授权公告日 2024.09.20

(21)申请号 202211141844.9
(22)申请日 2022.09.20
(65)同一申请的已公布的文献号
申请公布号 CN 115936064 A
(43)申请公布日 2023.04.07
(73)专利权人 电子科技大学
地址 611731 四川省成都市高新区(西区)西源大道2006号
(72)发明人 程筱舒 王忆文 娄鸿飞 李平
(51)Int.Cl.
G06N 3/0464(2023.01)
G06N 3/063(2023.01)
(56)对比文件
CN 107578098 A,2018.01.12

审查员 王佳佳

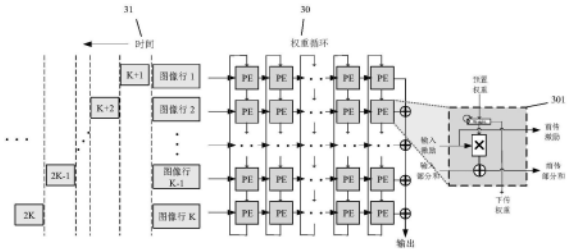
权利要求书1页 说明书3页 附图2页

(54)发明名称

一种基于权重循环数据流的神经网络加速阵列

(57)摘要

本发明具体涉及一种基于权重循环数据流的神经网络加速阵列,充分复用了从内存中读取的权重值和输入特征图数据,大大减少了对外部存储器的访问,属于神经网络的硬件加速技术领域。在人工智能芯片领域中,卷积运算占据整个卷积神经网络模型的计算量的百分之九十以上,本发明为了减少空域计算结构中,对输入数据的重复调用和移动,最大化数据复用,提出了权重循环数据流。通过设计一种基于权重循环数据流的PE阵列,对卷积操作进行优化,有效地降低硬件加速结构的功耗和延迟,从而提升系统的总体性能。



CN 115936064 B

1.一种基于权重循环数据流的神经网络加速阵列,其特征在于,PE阵列尺寸为卷积窗口的尺寸,PE单元为了数据移动而横向相连,为了权重循环移动而纵向循环互连;

将卷积核上的权重值预置到PE阵列中,在输入特征图数据首次充满PE阵列前,输入特征图不通过乘法器,而直接输出到右侧的PE单元,直到输入特征图数据充满PE阵列后再进行卷积操作;

卷积窗口的尺寸是K×K,首先同时输入图像行的1至K行,直到输入特征图数据首次充满PE阵列开始,按步长整体右移图像行,依次做卷积运算;当1至K行的输入特征图遍历完成后,更新一行图像行;其余未更新的K-1行数据继续重新循环输入;当图像行更新的时候,PE阵列的权重就按行整体循环下移,存储在每行的单个PE单元的权重寄存器中,PE阵列的K行图像行就整体右移做卷积;当图像行更新完时,PE阵列完成输入特征图最后K行的卷积后,运算结束;

在输入特征图右移过程中,权重在阵列的位置保持不变,数据每右移一位,阵列做一次卷积;

当输入特征图数据充满PE阵列,开始做卷积运算后,输入特征图与权重输入到乘法器中,产生一个部分和;该部分和与左侧来的输入部分和信号累加,产生前传部分和,输入到右侧的PE单元中;当每一行的部分和到达PE阵列的最右边时,由K-1个边缘加法器汇总,得到最终的一个卷积结果。

2.根据权利要求1所述的神经网络加速阵列,其特征在于,除了开始K行图像行同时进入外,图像行更新的间隔为PE阵列遍历卷积K行后的一个周期,而不是像脉动阵列每一拍就流动一次。

3.根据权利要求1所述的神经网络加速阵列,其特征在于,PE单元由至少一个数据寄存器、一个乘法器和一个加法器组成。

4.根据权利要求1所述的神经网络加速阵列,其特征在于,PE单元除了数据传输信号外,还有一些控制信号控制数据流动。

一种基于权重循环数据流的神经网络加速阵列

技术领域

[0001] 本发明涉及神经网络的硬件加速技术领域,具体涉及一种基于权重循环数据流的神经网络加速阵列设计方法。

背景技术

[0002] 随着物联网技术的飞速发展,可穿戴智能产品将人体各个部位作为互联网的接口,真正实现具有微型化、便携化、智能化可穿戴产品特点的人机一体化产品体验,为消费者提供便携式实时信息采集和数据服务,具有更大的技术含量和市场吸引力。

[0003] 不幸的是,设计高性能的可穿戴计算设备并非易事,其实现面临许多挑战。该领域是计算机科学和工程等不同研究领域的交叉点,使用了微电子和无线通信等各种技术。微电子技术的进步导致了适用于可穿戴计算设备的小尺寸低功耗人工智能芯片。

[0004] 在人工智能芯片领域中,隶属于机器学习范畴的深度学习,被广泛运用于图像分类、语音识别、对象检测等方面,并取得了显著的成果。卷积神经网络、递归神经网络和深度置信网络是深度学习研究的主要聚焦点,其中最先进的当属卷积神经网络。

[0005] 典型的卷积神经网络结构包括有:卷积层,激活层,池化层,全连接层和输入输出特征图等。其中卷积层的作用是特征提取,池化层的作用是像素压缩,全连接层的作用是分类。卷积层为计算密集型运算,而全连接层为数据密集型运算。

[0006] 卷积神经网络在数据处理方面有三个瓶颈问题:一是数据密集型,需要处理的数据量极大。二是计算密集型,对数据处理存储需要耗费大量计算资源和大量时间。三是速度失配问题,即数据处理速度慢于数据的产生速度。因此适合人工智能架构的专属芯片亟需发展,而高速数据传输、高速计算的神经网络加速器的实现将在多个方面有着重要的意义。

[0007] 目前,由于硬件性能的提高,在加速神经网络训练和推断过程中,主要采用CPU、GPU、FPGA和ASIC的形式。大约二十年前,CPU曾是实现神经网络算法的主流,其优化领域主要集中在软件部分。CNN不断增加的计算成本使得需要硬件加速其推理过程。在GPU方面,GPU集群可以并行地加速具有10亿多个参数的超大网络。主流的GPU聚类神经网络通常使用分布式SGD算法。许多研究进一步利用了这种并行性,努力实现不同集群之间的通信。由于FPGA具有许多吸引人的特性,因此成为CNN硬件加速的良好平台。一般来说,FPGA比CPU和GPU提供更高的能源效率,比CPU具有更高的性能。与GPU相比,FPGA的吞吐量是几十千兆次,内存访问有限。此外,它本身不支持浮点计算,但有更低能耗。专用集成电路ASIC是为特定应用而设计的专用处理器,虽然ASIC的灵活性较低,开发周期长,成本高,但是具有体积小、功耗低、计算速度快和可靠性高等优点。

[0008] 神经网络硬件加速器有两种较为典型的体系架构:时域计算架构(树状结构)和空域计算架构(PE阵列结构)。树状结构基于指令流对算数计算单元和存储资源进行集中控制,每个算数逻辑单元都从集中式存储系统获取运算数据,并向其写回结果。它由一个乘法加法树,一个用于分配输入值的缓冲区和一个预取缓冲区组成。PE阵列结构,每个算数运算单元都具有本地存储器,整个架构采用数据流控制,即所有的PE单元形成处理链关系,数据

直接在PE之间传递。它由全局缓冲区、FIFO和PE阵列组成。每个PE由一个或多个的乘法器和加法器组成,可实现高度并行计算。

[0009] 神经网络硬件加速器有四种较为典型的数据流模式:无局部复用数据流,输入固定流,输出固定流和权重固定流。对于无局部复用数据流,为了最大化存储容量和最小化片外存储器带宽,不给PE分配本地存储,而是把所有的区域分配给全局缓冲区以增加其容量,它必须多路传送输入特征图,单路传送卷积核权重,然后通过PE阵列累加部分和。对于输入固定流,计算核心把输入特征图读入局部的输入寄存器;计算核心充分复用这些输入数据,更新输出缓存中所有相关的输出部分和;更新后的输出部分和会重新写回输出缓存。对于输出固定流,计算核心把输入特征图的各通道读入局部的输入寄存器;存储在计算核心输出寄存器中的输出部分被充分复用,以完成三维卷积通道方向上的完全累加;最终的输出特征图会在池化之后再写入输出缓存。对于权重固定流,计算核心读取输入特征图分块到局部的输入寄存器;计算核心利用这些输入数据更新分块的输出部分和;存储在权重缓存中的分块卷积核权重被充分复用,以更新存储在输出缓存中的输出部分和。

[0010] 图1为一个简单的卷积层示意图,其中10为7×7的输入特征图,11为一个3×3的卷积核,12为5×5的输出特征图。卷积核窗口逐行呈“Z”字形在10上滑动做卷积运算,得到结果12。卷积运算占据整个卷积神经网络模型的计算量的百分之九十以上,因此通过设计一种结构化的PE阵列,对卷积操作进行优化,能有效地降低硬件加速结构的面积和功耗,从而提升系统的总体性能。

发明内容

[0011] 针对以上背景内容,本发明提出了一种针对神经网络卷积层运算的加速阵列设计方法。可以应用于FPGA或ASIC神经网络硬件加速器设计中,作为AI加速处理器的计算部分。

[0012] 在空域计算结构中,每个运算单元是通过数据流来进行控制的,因此关键就是解决数据流动问题。为了减少对输入数据的重复调用和移动,最大化数据复用,最好一次性输入特征图,因此提出了权重循环数据流(weight ring dataflow,WR)。

[0013] 本发明基于所提出的WR数据流,构建一种新的神经网络PE阵列架构。假设某一卷积层的卷积核尺寸为K^2,则相应PE阵列的尺寸也为K^2。PE单元为了数据移动而横向互连,为了权重循环而纵向循环互连。

[0014] 本发明的优点主要包括:充分复用了从内存中读取的权重值和输入特征图数据,大大减少了对外部存储器的访问,降低了整体功耗和延迟。

附图说明

[0015] 图1为卷积层示意图;

[0016] 图2为权重循环数据流示意图;

[0017] 图3为基于权重循环数据流的神经网络加速阵列示意图;

[0018] 图4为权重循环数据流PE阵列的工作流程图

具体实施方式

[0019] 以下结合附图对本发明的权重循环数据流和对应的阵列硬件实现进行详细说明:

[0020] 权重循环数据流具体体现如图2所示,20为尺寸为N^2的输入特征图,这里假设N取7,则第一个周期的卷积核201,尺寸为K^2,这里假设K取3,其滑动的行数范围为1至K行。在下一个周期的卷积核202,滑动的行数范围为2至K+1行,以此类推周期203的卷积核和204的卷积核等的情况;

[0021] 由图1卷积层的运算规律可知,11窗口内的数据在每次换行后,只更新一行数据。即原先的K行数据仅有一行被舍弃掉。因此若每次换行后都全部重新更新数据,会大大增加数据访存量。因此,将固有数据行保持位置的不变,仅用更新行覆盖舍弃的数据行即可。

[0022] 基于权重循环数据流的神经网络加速阵列如图3所示,含K^2个PE单元301的阵列是对卷积核窗口在图像上逐行滑动的模拟。首先将卷积核上的权重值预置到PE阵列30中,当K行图像行同时移入阵列做卷积的过程,原本称为权重固定流。但本研究提出的WR流稍许不同,它将每列的权重连接起来,提供了权值流动的通路,即30中的竖循环线。

[0023] 对于图4所示的PE阵列工作流程图,PE阵列的工作情况如下:

[0024] 1)首先同时输入图像行的1至K行,直到输入特征图数据首次充满30开始,按步长整体右移图像行,依次做卷积运算;

[0025] 2)当1至K行的输入特征图遍历完成后,更新一行图像行。其余未更新的K-1行数据继续重新循环输入;

[0026] 3)当31的每个周期图像行更新的时候,30的权重就按行整体循环下移,存储在每行的单个301的权重寄存器中,30的K行图像行就整体右移做卷积;

[0027] 4)当图像行更新完时,30完成输入特征图最后K行的卷积后,运算结束。

[0028] 单个301主要由乘法器、加法器和权重寄存器组成。预置权重首先输入到权重寄存器中存储起来,并不断地重复使用,直到更行图像行的时候,才分别循环下移到下一行的301的权重寄存器中。

[0029] 在输入特征图数据首次充满30前,输入激励不通过乘法器,直接输出到前传激励信号,将数据向右侧的301传递。

[0030] 当输入特征图数据充满30,开始做卷积运算后,输入激励与权重输入到乘法器中,产生一个部分和。该部分和与左侧来的输入部分和信号累加,产生前传部分和,输入到右侧的301中。当每一行的部分和到达30的最右边时,由K-1个边缘加法器汇总,得到最终的一个卷积结果。

[0031] 加法器部分除了上述结构外,还可替换成加法树的结构。

[0032] 以上所述,仅是本发明的较佳实施例而已,并非对本发明作任何形式上的限制。任何熟悉本领域的技术人员,在不脱离本发明技术方案范围情况下,都可利用上述提出的方法和技术内容对本发明技术方案做出一些可能的变动和修饰,或修改为等同变化的等效实施例。因此,凡是未脱离本发明技术方案的内容,依据本发明的技术实质对以上实施例所做的任何简单修改、等同变化及修饰,均仍属于本发明技术方案保护的范围内。

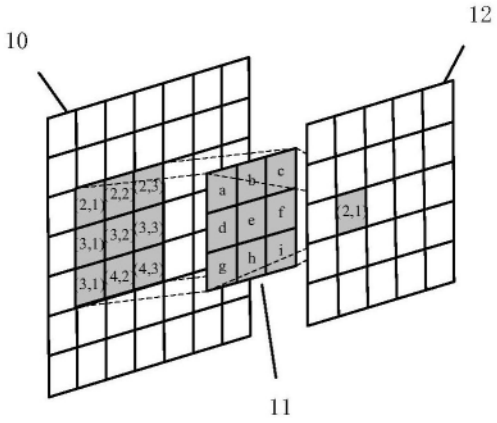

10
12
11
a
b
c
d
e
f
g
h
i
2,1
2,2
2,3
3,1
3,2
3,3
4,2
4,3

图1

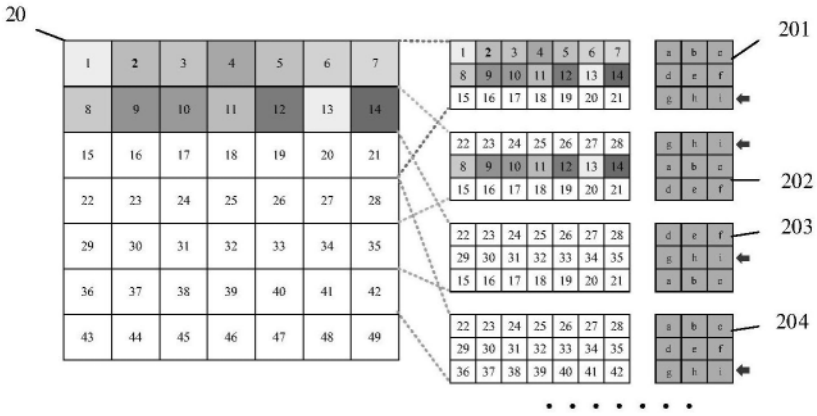

20
201
202
203
204

图2

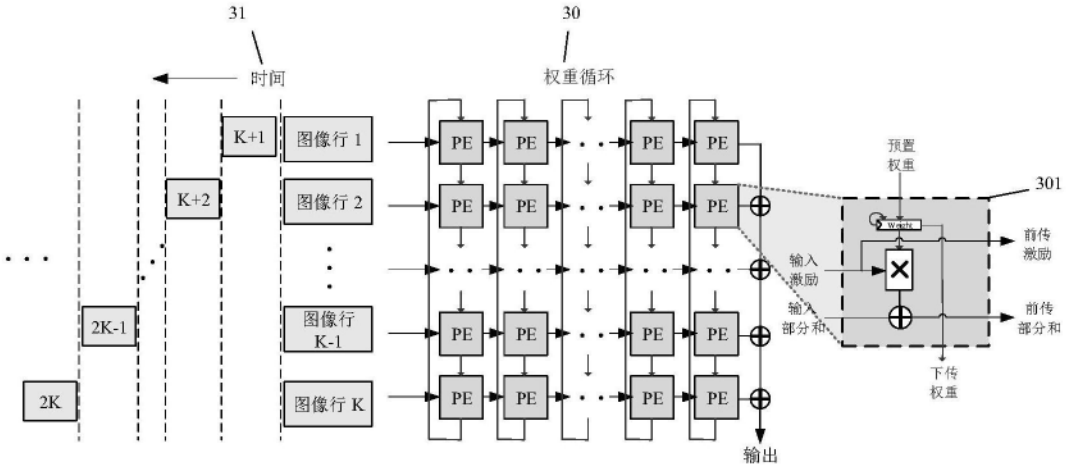

31
时间
30
权重循环
K+1
K+2
2K-1
2K
图像行 1
图像行 2
图像行 K-1
图像行 K
PE
预置权重
301
Weight
输入激励
输入部分和
前传激励
前传部分和
下传权重
输出

图3

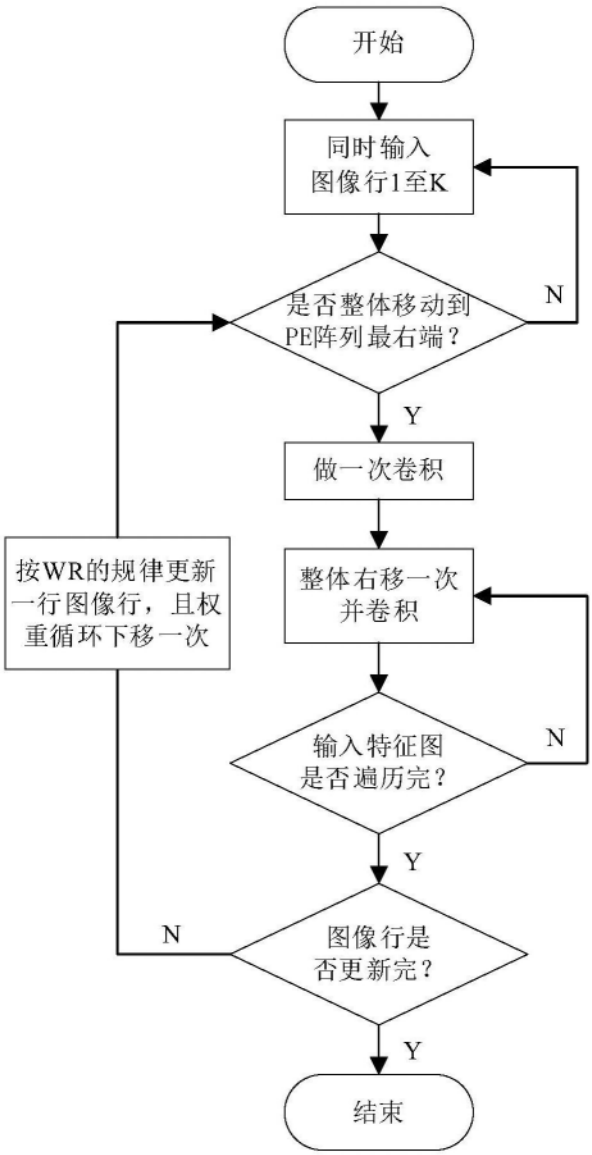

开始
同时输入
图像行1至K
是否整体移动到
PE阵列最右端？
N
Y
做一次卷积
整体右移一次
并卷积
输入特征图
是否遍历完？
N
Y
图像行是
否更新完？
N
Y
按WR的规律更新
一行图像行，且权
重循环下移一次
结束

图4