(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 111199274 A
(43)申请公布日 2020.05.26

(21)申请号 202010011113.7

(22)申请日 2020.01.06

(71)申请人 中科驭数(北京)科技有限公司
地址 100190 北京市海淀区科学院南路6号中国科学院计算技术研究所科研综合楼

(72)发明人 鄢贵海 卢文岩

(74)专利代理机构 北京金咨知识产权代理有限公司 11612
代理人 宋教花

(51)Int.Cl.
G06N 3/04(2006.01)
G06N 3/063(2006.01)

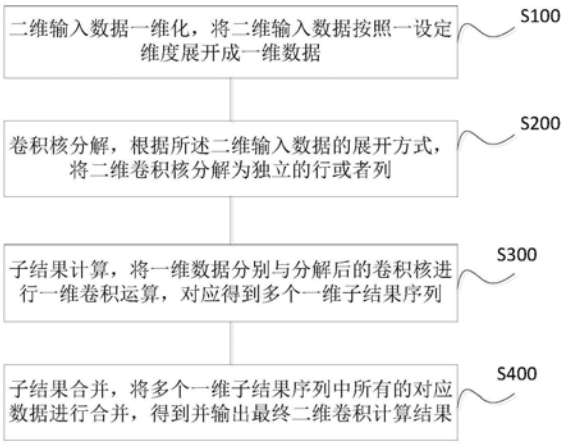
权利要求书2页 说明书6页 附图3页

(54)发明名称

二维卷积实现方法及装置

(57)摘要

本发明提供一种二维卷积实现方法及装置，该方法包括：二维输入数据一维化，将二维输入数据按照一设定维度展开成一维数据；卷积核分解，根据所述二维输入数据的展开方式，将二维卷积核分解为独立的行或者列；子结果计算，将所述一维数据分别与分解后的卷积核进行一维卷积运算，对应得到多个一维子结果序列；以及子结果合并，将所述多个一维子结果序列中所有的对应数据进行合并，得到并输出最终二维卷积计算结果。本发明还提供了二维卷积实现装置。本发明的方法和装置结合可以将所有的二维数据转化成一维数据，通过一维卷积运算即可完成二维卷积运算，大大简化了控制复杂度和片上数据路由难度。



CN 111199274 A

1.一种二维卷积实现方法,其特征在于,包括如下步骤:

二维输入数据一维化,将二维输入数据按照一设定维度展开成一维数据;

卷积核分解,根据所述二维输入数据的展开方式,将二维卷积核分解为独立的行或者列;

子结果计算,将所述一维数据分别与分解后的卷积核进行一维卷积运算,对应得到多个一维子结果序列;以及

子结果合并,将所述多个一维子结果序列中所有的对应数据进行合并,得到并输出最终二维卷积计算结果。

2.如权利要求1所述的二维卷积实现方法,其特征在于,将所述二维输入数据按照行、列、对角线或不规则的方式展开成所述一维数据。

3.如权利要求2所述的二维卷积实现方法,其特征在于,所述卷积核分解的方式与将所述二维输入数据展开成所述一维数据的方式完全一致。

4.如权利要求1、2或3所述的二维卷积实现方法,其特征在于,

所述二维输入数据一维化进一步包括:将所述二维输入数据按行的方式展开,相邻行尾首相连成所述一维数据;

所述卷积核分解进一步包括:将所述二维卷积核同样以行的形式分解成卷积核第一行和卷积核第二行;

所述子结果计算进一步包括:所述一维数据分别与所述卷积核第一行和卷积核第二行进行一维卷积运算,所述一维数据与所述卷积核第一行进行一维卷积运算时,滑动窗从所述一维数据的第一行开始移动到倒数第二行结束;所述一维数据与所述卷积核第二行进行一维卷积运算时,滑动窗从所述一维数据的第二行开始移动到最后一行结束;以及

所述子结果合并进一步包括:所述一维数据与所述卷积核第一行及卷积核第二行进行一维卷积运算的子结果分别对应相加合并,得到最终计算结果。

5.如权利要求4所述的二维卷积实现方法,其特征在于,所述一维化输入数据与分解后的卷积核在一维卷积运算时产生的中间卷积结果,在后级子结果合并运算时与之前计算所得的子结果进行累加并缓存,然后与新计算所得的子结果进行累加合并,直至与所有分解后的卷积核相关的子结果计算并合并完成。

6.一种二维卷积实现装置,其特征在于,包括:

一维卷积运算单元,用于实现通过将二维输入数据按照一设定维度展开后得到的一维化的输入数据与通过将二维卷积核按照所述二维输入数据的展开方式分解后得到的卷积核之间的一维卷积运算,得到多个一维子结果序列;

子结果合并运算单元,用于完成多个一维子结果的合并运算;以及

控制器,分别与所述一维卷积运算单元和子结果合并运算单元连接,用于产生控制信号,以协调所述一维卷积运算单元和子结果合并运算单元之间的运行。

7.如权利要求6所述的二维卷积实现装置,其特征在于,所述一维卷积运算单元为乘累加树形式,所述控制器产生相应控制信号,用于实现一维化输入数据卷积的不连续控制。

8.如权利要求6或7所述的二维卷积实现装置,其特征在于,还包括多个缓冲单元,所述多个缓冲单元用于同步各数据流,所述多个缓冲单元进一步包括:

输入数据缓冲单元,用于缓存所述一维化输入数据,分别与所述控制器和所述一维卷

积运算单元连接；

卷积核缓冲单元，用于缓存分解后的卷积核数据，分别与所述控制器和所述一维卷积运算单元连接；以及

子结果缓冲单元，用于缓存所述一维卷积运算单元的计算结果，分别与所述控制器、一维卷积运算单元及所述子结果合并运算单元连接。

9.如权利要求8所述的二维卷积实现装置，其特征在于，所述多个缓冲单元还包括：

合并结果缓冲单元，用于重新读入并缓存上次的一维化输入数据与卷积核第一行的卷积结果，分别与所述控制器和所述子结果合并运算单元连接。

10.如权利要求9所述的二维卷积实现装置，其特征在于，所述输入数据缓冲单元重新读入所述一维化输入数据并缓存；所述卷积核缓冲单元读入所述卷积核第一行并缓存；合并结果缓冲单元重新读入上次的所述一维化输入数据与所述卷积核第一行的中间卷积结果并缓存；所述一维卷积运算单元分别从所述输入数据缓冲单元和卷积核缓冲单元中读取所述一维化输入数据和卷积核第二行的数据进行一维卷积运算，并将运算结果缓存到所述子结果缓冲单元中；所述子结果合并运算单元从所述子结果缓冲单元读取本次一维卷积运算单元计算输出的子结果，从所述合并结果缓冲单元读取所述中间卷积结果，并进行累加合并后输出最终计算结果，直至完成最终的二维卷积运算。

二维卷积实现方法及装置

技术领域

[0001] 本发明涉及数据及数据库处理，特别是一种排序数据处理的二维卷积实现方法及装置。

背景技术

[0002] 二维卷积是众多科学计算中最核心的操作之一，占据了整个应用中绝大部分的计算量，如卷积神经网络各层中输入特征图像与二维卷积核之间的计算，占据了整个网络模型中超90％的计算量。因此，提高二维卷积的计算效率，是众多核心科学计算优化的关键。

[0003] 为了加速二维卷积的计算效率，越来越多的专用计算架构被提出，该类计算架构多根据二维卷积的特点而设计，计算单元多以二维的形式组织，直接处理二维数据，计算效率较高。但是，二维数据寻址方式非常复杂，给芯片设计带来了巨大的挑战。为了实现如此复杂的寻址，及时地为计算单元提供数据，现有方案多利用多层片上存储结构，并精细对存储结构进行划分和数据排列，控制非常复杂。现有设计，虽然采用了如此复杂的方案，仍经常出现数据冲突导致计算单元数据不足的问题，造成整体计算性能低下。

发明内容

[0004] 本发明所要解决的技术问题是针对现有技术二维卷积寻址复杂、控制复杂的问题，提供一种二维卷积实现方法及装置。

[0005] 为了实现上述目的，本发明提供了一种二维卷积实现方法，其中，包括如下步骤：

[0006] 二维输入数据一维化，将二维输入数据按照一设定维度展开成一维数据；

[0007] 卷积核分解，根据所述二维输入数据的展开方式，将二维卷积核分解为独立的行或者列；

[0008] 子结果计算，将所述一维数据分别与分解后的卷积核进行一维卷积运算，对应得到多个一维子结果序列；以及

[0009] 子结果合并，将所述多个一维子结果序列中所有的对应数据进行合并，得到并输出最终二维卷积计算结果。

[0010] 可选地，上述的二维卷积实现方法中，将所述二维输入数据按照行、列、对角线或不规则的方式展开成所述一维数据。

[0011] 可选地，上述的二维卷积实现方法中，所述卷积核分解的方式与将所述二维输入数据展开成所述一维数据的方式完全一致。

[0012] 可选地，上述的二维卷积实现方法中，所述二维输入数据一维化进一步包括：将所述二维输入数据按行的方式展开，相邻行尾首相连成所述一维数据；

[0013] 所述卷积核分解进一步包括：将所述二维卷积核同样以行的形式分解成卷积核第一行和卷积核第二行；

[0014] 所述子结果计算进一步包括：所述一维数据分别与所述卷积核第一行和卷积核第二行进行一维卷积运算，所述一维数据与所述卷积核第一行进行一维卷积运算时，滑动窗

从所述一维数据的第一行开始移动到倒数第二行结束;所述一维数据与所述卷积核第二行进行一维卷积运算时,滑动窗从所述一维数据的第二行开始移动到最后一行结束;以及

[0015] 所述子结果合并进一步包括:所述一维数据与所述卷积核第一行及卷积核第二行进行一维卷积运算的子结果分别对应相加合并,得到最终计算结果。

[0016] 可选地,上述的二维卷积实现方法中,所述一维化输入数据与分解后的卷积核在一维卷积运算时产生的中间卷积结果,在后级子结果合并运算时与之前计算所得的子结果进行累加并缓存,然后与新计算所得的子结果进行累加合并,直至与所有分解后的卷积核相关的子结果计算并合并完成。

[0017] 为了更好地实现上述目的,本发明还提供了一种二维卷积实现装置,该装置包括:

[0018] 一维卷积运算单元,用于实现通过将二维输入数据按照一设定维度展开后得到的一维化的输入数据与通过将二维卷积核按照所述二维输入数据的展开方式分解后得到的卷积核之间的一维卷积运算,得到多个一维子结果序列;

[0019] 子结果合并运算单元,用于完成多个一维子结果的合并运算;以及

[0020] 控制器,分别与所述一维卷积运算单元和子结果合并运算单元连接,用于产生控制信号,以协调所述一维卷积运算单元和子结果合并运算单元之间的运行。

[0021] 上述的二维卷积实现装置,其中,所述一维卷积运算单元为乘累加树形式,所述控制器产生相应控制信号,用于实现一维化输入数据卷积的不连续控制。

[0022] 上述的二维卷积实现装置,其中,还包括多个缓冲单元,所述多个缓冲单元用于同步各数据流,所述多个缓冲单元进一步包括:

[0023] 输入数据缓冲单元,用于缓存所述一维化输入数据,分别与所述控制器和所述一维卷积运算单元连接;

[0024] 卷积核缓冲单元,用于缓存分解后的卷积核数据,分别与所述控制器和所述一维卷积运算单元连接;以及

[0025] 子结果缓冲单元,用于缓存所述一维卷积运算单元的计算结果,分别与所述控制器、一维卷积运算单元及所述子结果合并运算单元连接。

[0026] 上述的二维卷积实现装置,其中,所述多个缓冲单元还包括:

[0027] 合并结果缓冲单元,用于重新读入并缓存上次的一维化输入数据与卷积核第一行的卷积结果,分别与所述控制器和所述子结果合并运算单元连接。

[0028] 上述的二维卷积实现装置,其中,所述输入数据缓冲单元重新读入所述一维化输入数据并缓存;所述卷积核缓冲单元读入所述卷积核第一行并缓存;合并结果缓冲单元重新读入上次的所述一维化输入数据与所述卷积核第一行的中间卷积结果并缓存;所述一维卷积运算单元分别从所述输入数据缓冲单元和卷积核缓冲单元中读取所述一维化输入数据和卷积核第二行的数据进行一维卷积运算,并将运算结果缓存到所述子结果缓冲单元中;所述子结果合并运算单元从所述子结果缓冲单元读取本次一维卷积运算单元计算输出的子结果,从所述合并结果缓冲单元读取所述中间卷积结果,并进行累加合并后输出最终计算结果,直至完成最终的二维卷积运算。

[0029] 本发明实施例的技术效果包括:

[0030] 本发明实施例的二维卷积一维化的方法,克服了二维卷积操作寻址复杂、控制难度大的问题。可以将所有的二维数据转化成一维数据,通过一维卷积运算即可完成二维卷

积运算，大大降低了片上数据存储难度，简化了路由复杂度及控制复杂度。同时，本发明实施例的装置可以完美地支持一维化卷积运算，大大提高了计算效率，在不降低处理性能的前提下，可以降低数据在存储中的寻址难度，以及整个计算控制的复杂度。

[0031] 本领域技术人员将会理解的是，能够用本发明实现的目的和优点不限于以上具体所述，并且根据以下详细说明将更清楚地理解本发明能够实现的上述和其他目的。

附图说明

[0032] 图1为本发明一实施例的二维卷积实现方法流程示意图；

[0033] 图2为本发明一实施例的二维卷积实现装置结构示意图；

[0034] 图3为本发明一实施例的一维卷积运算单元示意图；

[0035] 图4为一个二维卷积的示例；

[0036] 图5为本发明一实施例的二维卷积一维化过程示意图。

具体实施方式

[0037] 为使本发明的目的、技术方案和优点更加清楚明白，下面结合实施方式和附图，对本发明做进一步详细说明。在此，本发明的示意性实施方式及其说明用于解释本发明，但并不作为对本发明的限定。

[0038] 在此，还需要说明的是，为了避免因不必要的细节而模糊了本发明，在附图中仅仅示出了与根据本发明的方案密切相关的结构和/或处理步骤，而省略了与本发明关系不大的其他细节。

[0039] 应该强调，术语“包括/包含/具有”在本文使用时指特征、要素、步骤或组件的存在，但并不排除一个或更多个其它特征、要素、步骤或组件的存在或附加。

[0040] 在此，还需要说明的是，在不冲突的情况下，本申请中的实施例及实施例中的特征可以相互结合。

[0041] 为了克服二维卷积操作寻址复杂、控制难度大的问题，本发明实施例提出了一种二维卷积操作一维化的方法。图1为本发明一实施例的二维卷积实现方法示意图，参见图1，本实施例的二维卷积实现方法包括如下步骤：

[0042] 步骤S100，二维输入数据一维化，将二维输入数据按照一设定维度展开成一维数据。

[0043] 步骤S200，卷积核分解，根据步骤S100二维输入数据的展开方式，将二维卷积核分解为独立的行或者列。

[0044] 步骤S300，子结果计算，将步骤S100所述一维数据分别与步骤S200分解后的卷积核进行一维卷积运算，对应得到多个一维子结果序列。

[0045] 步骤S400，子结果合并，将步骤S300计算所得所述多个一维子结果序列中所有的对应数据进行合并，得到并输出最终二维卷积计算结果。

[0046] 在步骤S100中，设定维度可以是按照行的维度或按照列的维度，也可以是按照对角线维度的方式，也即二维输入数据的展开方式可以是将二维输入数据按行和按列展开成一维序列，但也可以按照对角线的方式，或其他更加不规则的方式展开成一维数据。

[0047] 在步骤S200中，卷积核分解的方式与步骤S100中将二维输入数据展开成一维数据

的方式相同，例如，如果输入数据按行展开，卷积核也按行进行分解。

[0048] 在步骤S300中，一维化的输入数据要分别与步骤S200中卷积核分解的各部分进行一维卷积运算，分别得到多个一维的子结果。

[0049] 在步骤S400中，所有一维子结果中对应元素进行求和操作得到最终二维卷积计算结果。

[0050] 参见图2，图2为本发明一实施例中用于快速实现二维卷积的二维卷积实现装置结构示意图。本实施例的二维卷积实现装置包括：一维卷积运算单元、子结果合并运算单元以及控制器。

[0051] 一维卷积运算单元用于实现一维化的输入数据与分解后的卷积核之间的一维卷积运算。一维卷积运算单元的组织结构相对灵活，任何可以实现一维卷积操作的单元都可以，最典型的结构是乘累加树的形式，但不限于此形式。但需要控制器的额外控制信号，来实现一维化输入数据卷积的不连续控制，如按行扩展的形式，每行结束切换到下一行数据时需要开始新的卷积窗口计算。

[0052] 子结果合并运算单元用于完成多个子结果的合并运算。

[0053] 控制器分别与一维卷积运算单元和子结果合并运算单元连接，用于产生控制信号，以协调一维卷积运算单元和子结果合并运算单元之间的运行。

[0054] 本实施例中，二维卷积实现装置还可包括多个缓冲单元，该多个缓冲单元用于同步各数据流，所述多个缓冲单元进一步包括：输入数据缓冲单元、卷积核缓冲单元、子结果缓冲单元和合并结果缓冲单元。

[0055] 输入数据缓冲单元用于缓存一维化输入数据，分别与控制器和一维卷积运算单元连接。

[0056] 卷积核缓冲单元用于缓存分解后的卷积核数据，分别与所述控制器和一维卷积运算单元连接。

[0057] 子结果缓冲单元用于缓存一维卷积运算单元的中间计算结果，分别与控制器、一维卷积运算单元及所述子结果合并运算单元连接。

[0058] 合并结果缓冲单元用于重新读入并缓存上次的一维化输入数据与卷积核第一行的卷积结果，分别与所述控制器和所述子结果合并运算单元连接。

[0059] 一维化输入数据与部分分解的卷积核在一维卷积运算单元产生的中间卷积结果，会在后级子结果合并运算单元与之前计算所得的子结果进行累加，合并后的子结果，在一维化的输入数据与另一部分分解的卷积核进行一维卷积运算的时候，重新被输入回该计算架构中，并通过合并结果缓冲单元进行缓存，然后在子结果合并运算单元中与新计算所得子结果进行累加合并。依次类推，直到与所有分解的卷积核相关的子结果被计算并合并完成，整个二维卷积计算完成。其中输入数据缓冲单元、卷积核缓冲单元、合并结果缓冲单元、子结果缓冲单元是为了同步各数据流，在数据本身同步性比较好的情况下，上述各缓冲单元不是非必要模块，可以省略。

[0060] 参见图3，图3为本发明一实施例的一维卷积运算单元示意图。本实施例中，一维卷积运算单元为乘累加树形式，控制器产生相应控制信号，用于实现一维化输入数据卷积的不连续控制。

[0061] 一维卷积运算的工作过程如下：

[0062] 首先一维化的输入数据和卷积核第一行分别输入到输入数据缓冲单元和卷积核缓冲单元进行缓存，然后一维卷积运算单元分别从输入数据缓冲单元逐个取出输入数据，从卷积核缓冲单元取出卷积核第一行各个数据元素，进行一维卷积运算，其计算结果在子结果缓冲单元中进行缓存。由于，这是一维化输入数据与卷积核第一行进行的一维卷积运算，之间没有输出的计算结果需要进行合并，所以其结果可以直接输出进行存储，一直到整个一维输入数据结束。

[0063] 之后，将一维化的输入数据重新读入到输入数据缓冲单元进行缓存，将卷积核第一行读入到卷积核缓冲单元进行缓存，同时将上次一维化输入结果与卷积核第一行的卷积结果重新读入缓存到合并结果缓冲单元中。

[0064] 接着，一维卷积运算单元分别从输入数据缓冲单元中和卷积核缓冲单元中读取一维化输入数据以及卷积核第二行数据进行一维卷积运算，并将运算结果缓存到子结果缓冲单元中。

[0065] 最后，子结果合并运算单元从子结果缓冲单元读取本次一维卷积运算单元计算输出的中间卷积结果，从合并结果缓冲单元中读取之间结果的中间卷积结果，进行累加合并，并将最终计算结果输出，一直到一维化输入数据结果完成最终的二维卷积运算。

[0066] 图4所示为一个二维卷积实现示例。输入二维数据与二维卷积核以滑动窗的形式进行卷积运算，滑动窗口可分别按行按列两个方向移动一直到行和列结尾，滑动窗内数据元素与卷积核内对应元素一一相乘并将计算结果累加完成一个计算结果。

[0067] 图5为本发明一实施例的将二维卷积进行一维化的过程示意图。本实施例中，二维输入数据一维化的过程进一步包括：

[0068] 首先，将二维输入数据按行的方式展开，相邻行尾首相连成一维化的输入数据(简称一维数据)；第二，对卷积核进行分解：将二维卷积核同样以行的形式分解成卷积核第一行和卷积核第二行；第三，进行子结果计算，进一步包括：一维数据分别与卷积核第一行和卷积核第二行进行一维卷积运算，一维数据与卷积核第一行进行一维卷积运算时，滑动窗从一维数据的第一行开始移动到倒数第二行结束；一维数据与卷积核第二行进行一维卷积运算时，滑动窗从一维数据的第二行开始移动到最后一行结束。最后进行子结果合并，进一步包括：一维数据与卷积核第一行及卷积核第二行进行一维卷积运算的子结果分别对应相加合并，得到最终计算结果。

[0069] 其中，一维化输入数据与分解后的卷积核在一维卷积运算时产生的中间卷积结果，在后级子结果合并运算时与之前计算所得的子结果进行累加并缓存，然后与新计算所得的子结果进行累加合并，直至与所有分解后的卷积核相关的子结果计算并合并完成。

[0070] 在卷积操作中，很多操作是类累加操作。很多时候无法一次完成所有操作得到最终结果，因此，需要多次迭代累加才能完成操作，即将前次合并后子结果作为输入与本次计算结果合并得到中间结果，或最终结果。

[0071] 本发明实施例的方法可以将所有的二维数据转化成一维数据，通过一维卷积运算即可完成二维卷积运算，大大简化控制复杂度和片上数据路由难度。同时，与本发明实施例的装置结合，大大提高了计算效率，在不降低处理性能的前提下，降低了数据在存储中的寻址难度，以及整个计算控制复杂度。

[0072] 本领域普通技术人员应该可以明白，结合本文中所公开的实施方式描述的各示例

性的组成部分、系统和方法，能够以硬件、软件或者二者的结合来实现。具体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。当以硬件方式实现时，其可以例如是电子电路、专用集成电路(ASIC)、适当的固件、插件、功能卡等等。当以软件方式实现时，本发明的元素是被用于执行所需任务的程序或者代码段。程序或者代码段可以存储在机器可读介质中，或者通过载波中携带的数据信号在传输介质或者通信链路上传送。“机器可读介质”可以包括能够存储或传输信息的任何介质。机器可读介质的例子包括电子电路、半导体存储器设备、ROM、闪存、可擦除ROM(EROM)、软盘、CD-ROM、光盘、硬盘、光纤介质、射频(RF)链路，等等。代码段可以经由诸如因特网、内联网等的计算机网络被下载。

[0073] 还需要说明的是，本发明中提及的示例性实施例，基于一系列的步骤或者装置描述一些方法或系统。但是，本发明不局限于上述步骤的顺序，也就是说，可以按照实施例中提及的顺序执行步骤，也可以不同于实施例中的顺序，或者若干步骤同时执行。

[0074] 本发明中，针对一个实施方式描述和/或例示的特征，可以在一个或更多个其它实施方式中以相同方式或以类似方式使用，和/或与其他实施方式的特征相结合或代替其他实施方式的特征。

[0075] 以上所述仅为本发明的优选实施例而已，并不用于限制本发明，对于本领域的技术人员来说，本发明实施例可以有各种更改和变化。凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

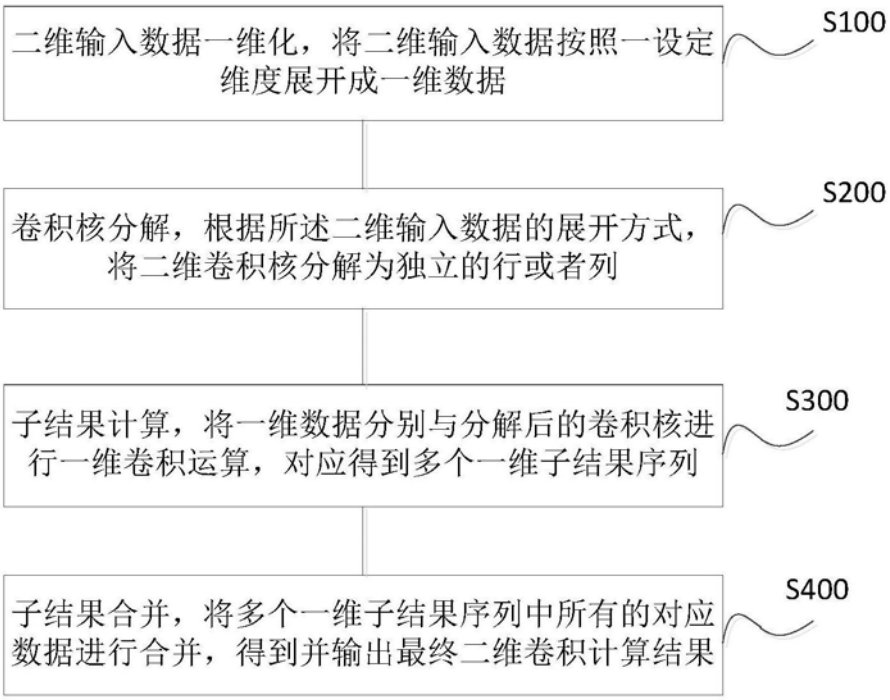

二维输入数据一维化，将二维输入数据按照一设定维度展开成一维数据
S100
卷积核分解，根据所述二维输入数据的展开方式，将二维卷积核分解为独立的行或者列
S200
子结果计算，将一维数据分别与分解后的卷积核进行一维卷积运算，对应得到多个一维子结果序列
S300
子结果合并，将多个一维子结果序列中所有的对应数据进行合并，得到并输出最终二维卷积计算结果
S400

图1

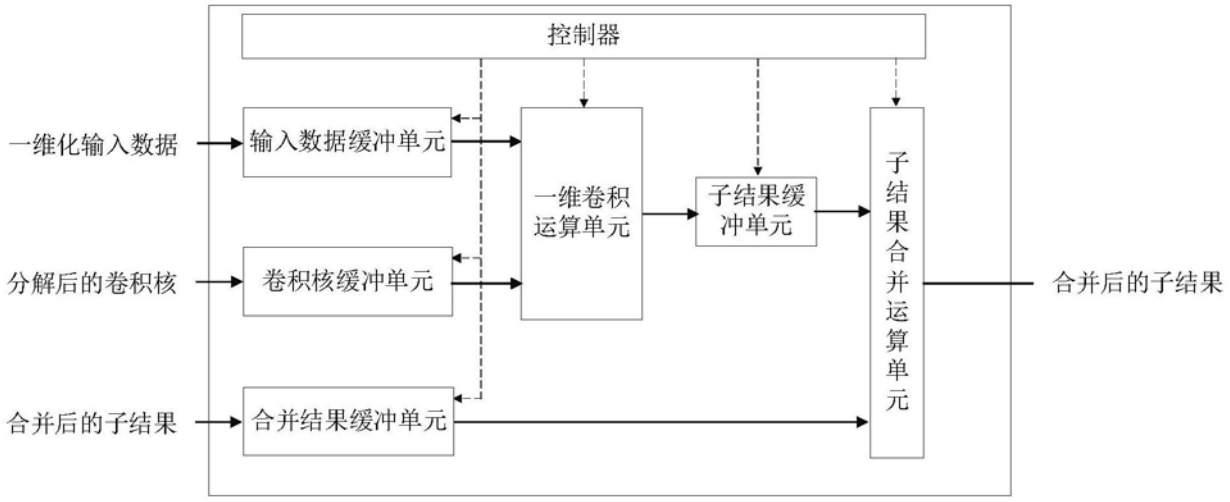

控制器
一维化输入数据
输入数据缓冲单元
分解后的卷积核
卷积核缓冲单元
合并后的子结果
合并结果缓冲单元
一维卷积运算单元
子结果缓冲单元
子结果合并运算单元
合并后的子结果

图2

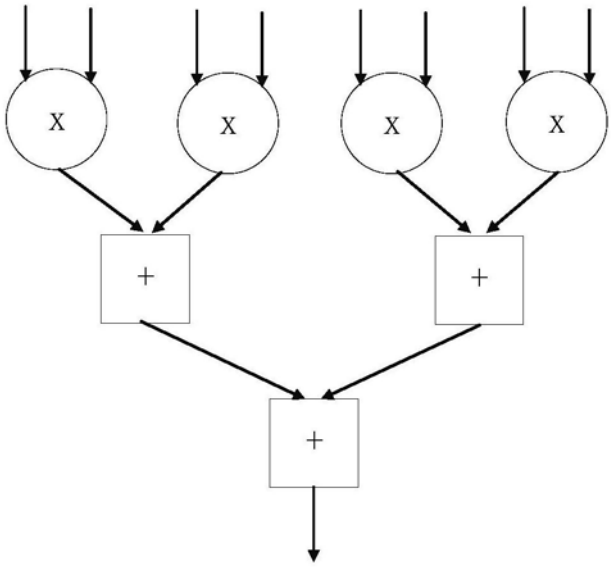


图3

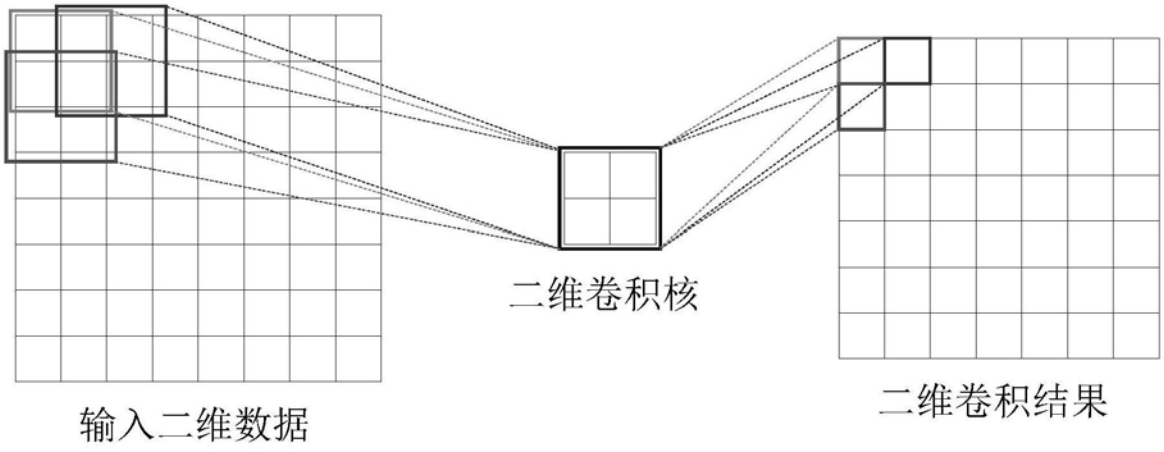


图4

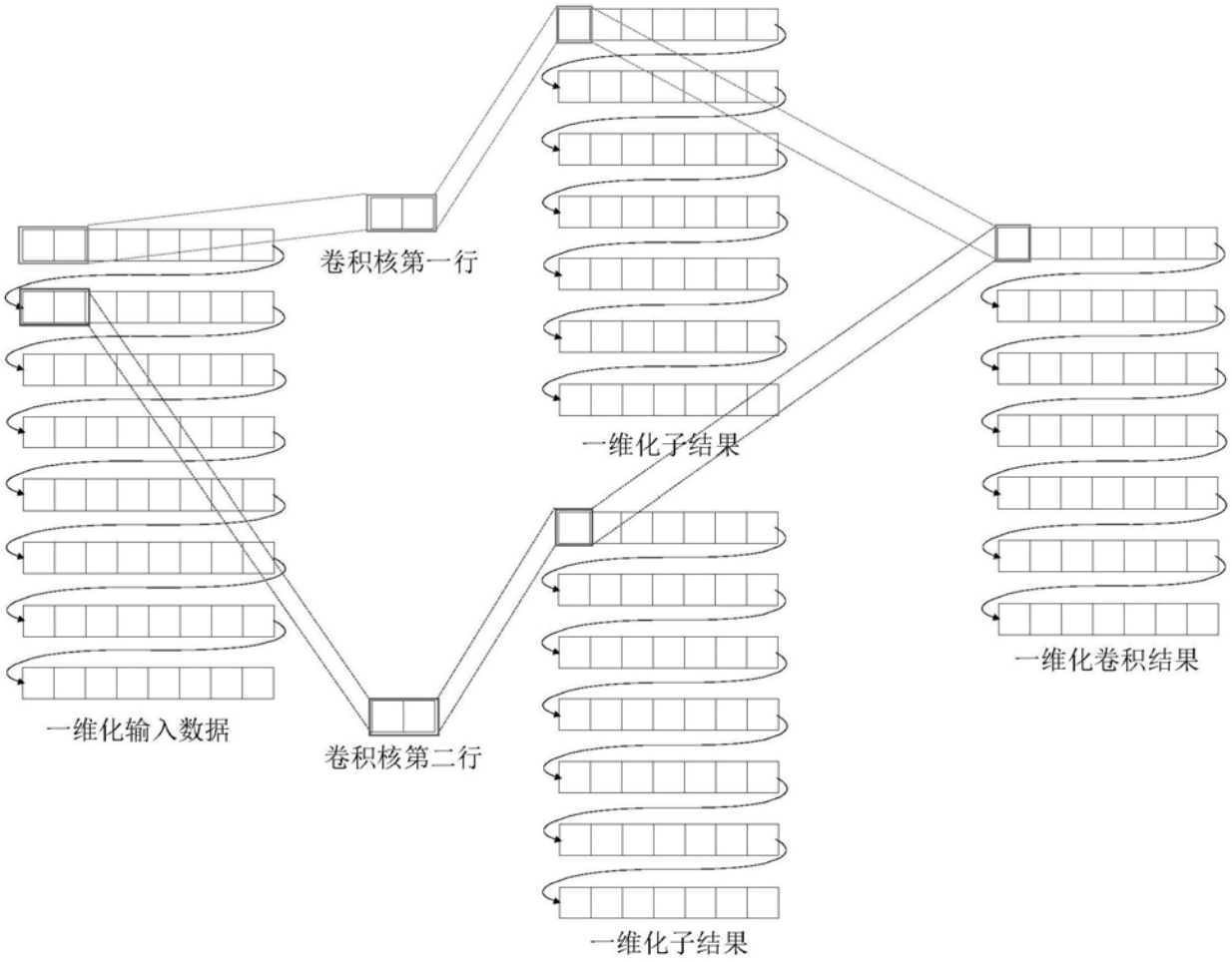


图5