(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 106779060 A
(43)申请公布日 2017.05.31

(21)申请号 201710071825.6

(22)申请日 2017.02.09

(71)申请人 武汉魅瞳科技有限公司
地址 430074 湖北省武汉市洪山区珞喻路1037号华中科技大学老保卫处101

(72)发明人 李开 邹复好 章国良 黄浩 杨帆 孙浩

(74)专利代理机构 武汉东喻专利代理事务所(普通合伙) 42224
代理人 张英

(51)Int.Cl.
*G06N 3/063*(2006.01)
*G06N 3/08*(2006.01)

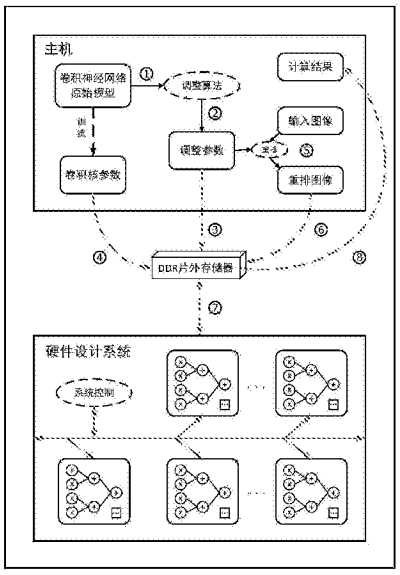
权利要求书2页 说明书15页 附图9页

(54)发明名称

一种适于硬件设计实现的深度卷积神经网络的计算方法

(57)摘要

本发明提出了一种适于硬件设计实现的深度卷积神经网络的计算方法，该计算方法提出预先利用相关调整参数重新调整深度卷积神经网络的计算结构，打破传统卷积神经网络中计算窗口结构固定的束缚，使每一计算层先参与计算的数据能够先到达，充分挖掘出深度卷积神经网络中的计算并行性以及各计算层之间的流水性，以有效地减少大量中间结果的存储。按照本发明提出的方法调整后的深度卷积网络计算结构更有利于在专用硬件设计上高效并行流水化实现，并且有效地解决了计算过程中由于各类填充操作而导致的资源浪费和有效计算延误问题，能有效地降低系统功耗和大大提高运算处理速度。




CN 106779060 A

1.一种适于硬件设计实现的深度卷积神经网络的计算方法，其特征在于，该计算方法包括如下步骤：

STEP1：对于给定的深度卷积神经网络模型，在上层主机中利用深度卷积神经网络计算结构调整算法，预先生成与该深度卷积神经网络模型相对应的特定调整参数；

STEP2：将所述特定调整参数、训练好的卷积核参数从所述上层主机上加载置DDR片外存储器中；

STEP3：硬件设计系统从所述外存储器中直接加载全部的所述特定调整参数，完成所述特定调整参数的部署，利用所述特定调整参数调整深度卷积神经网络的计算结构；

STEP4：所述上层主机结合所述特定调整参数中提供的原始输入图像位置重排参数对输入图像进行像素点位置重排，并将重排后的图像加载至所述存储器中，接着启动深度卷积神经网络的计算过程；

STEP5：所述硬件设计系统不断从所述DDR片外存储器中获取计算输入数据，在所述特定调整参数和所述卷积核参数的协同参与下完成相关计算过程。

2.如权利要求1所述的适于硬件设计实现的深度卷积神经网络的计算方法，其特征在于：

所述特定调整参数分为两类：计算顺序序列化参数和填充过滤参数；所述计算顺序序列化参数为原始输入图像位置重排参数、新旧值选取标记参数和旧值选取地址参数；

所述填充过滤参数为核参地址跳跃标记参数、跳跃目的核参地址参数和窗口计算提前结束标记参数；

在深度卷积神经网络的池化层中，所述填充过滤参数单指窗口计算提前结束标记参数；

其中，所述原始输入图像位置重排参数用于对所述上层主机中的输入图像进行像素点位置重排以得到重排后的图像；

所述新旧值选取标记参数为其所在层的计算顺序序列化实现过程提供新旧值数据选取标记值指定，所述标记值指定是从上一层的特征图中顺序获取下一个参与计算的新值数据还是从已经获取过的新值数据中选取旧值数据，当所述新旧值选取标记参数指定从已经获取过的新值数据中选取旧值数据时，所述旧值选取地址参数为其提供选取旧值数据的地址；

其中，所述核参地址跳跃标记参数在深度卷积神经网络的卷积层中指明当前计算位置之后是否存在填充元素，当存在填充元素时，则需要执行跳跃过滤操作，所述跳跃目的核参地址参数为其提供卷积核参数的跳跃目的地址，当一个原始计算窗口中存在填充元素时，由于跳跃过滤操作的存在，计算窗口中真正投入计算的元素数量将小于原始计算窗口大小，此时，所述窗口计算提前结束标记参数为其提供窗口计算提前结束标记。

3.如权利要求2所述的适于硬件设计实现的深度卷积神经网络的计算方法，其特征在于：

所述STEP1中涉及的深度卷积神经网络计算结构调整算法，采用队列为主要数据结构遍历各层，以首层全连接层为起点，以起始输入图像层为终点，遍历过程中生成与每层相关的所述调整参数，每一层中的所有特征图在后续参与计算时共享与该层对应的一套所述特定调整参数；

其中，所述结构调整算法的具体步骤如下：

STEP1-1以首层全连接层中所输入的单张特征图的元素排列顺序为初始排列顺序，并将表示该初始排列顺序的一维位置序号序列依次存入至队列中；

STEP1-2判断所述队列是否为空，为空时算法结束，否则转至下一步骤STEP1-3；

STEP1-3每次取队列队首位置序号进行扩充，根据所在层的神经元结构找到与该位置序号所在元素相对应的上层特征图中的计算窗口位置，并依次分析该计算窗口中的每个元素在其所在的单张特征图中的位置；

STEP1-4判断当前窗口是否分析完毕，若没有分析完毕，则转至步骤STEP1-5，否则，转至步骤STEP1-10；

STEP1-5分析下一个当前窗口中的元素，判断该元素是否处于所在特征图中的填充位置，若否，转至STEP1-6；否则转至STEP1-9；

STEP1-6为该层中的此次分析行为分配一个唯一的有效分析序号，所述有效分析序号从编号1开始依次递增分配，并判断其所对应位置的元素于其所在的单张特征图中是否被首次分析到，若是，转至STEP1-7；否则转至STEP1-8；

STEP1-7将当前有效分析序号的新旧值选取标记值置为1，其中标记值为1表示选取新值；标记值为0表示选取旧值，并判断有效分析序号所对应位置的元素于是否处于起始输入图像层，若是，将当前有效分析序号添加到所述原始输入图像位置重排参数中；否则，将当前有效分析序号添加到所述队列队尾，转至所述STEP1-4；

STEP1-8将当前有效分析序号的新旧值选取标记值置为0，转至所述STEP1-4；

STEP1-9为该层中的此次分析行为分配一个唯一的无效分析序号，所述无效分析序号从编号1开始依次递增分配，并判断该无效分析序号是否位于一段连续无效分析序号的段首，若是，将其正前面的一个有效分析序号添加至所述核参地址跳跃标记参数中，将紧接在该段连续无效分析序号末尾的一个有效分析序号添加至所述跳跃目的核参地址参数中，转至所述STEP1-4；否则，直接转至所述STEP1-4。

STEP1-10判断分析完的计算窗口中是否出现过处于填充位置的元素，若是，将该计算窗口中最后一个有效分析序号添加至窗口计算提前结束标记参数中，转至所述STEP1-2；否则，直接转至所述STEP1-2。

# 一种适于硬件设计实现的深度卷积神经网络的计算方法

## 技术领域

[0001] 本发明属于复杂算法加速方法，具体涉及一种适于硬件设计实现的深度卷积神经网络的计算方法。

## 背景技术

[0002] 伴随着深度学习掀起的新的机器学习热潮，深度卷积神经网络已经广泛应用于语音识别、图像识别和自然语音处理等不同的大规模机器学习问题中，并取得了一系列突破性的研究成果，其强大的特征学习与分类能力引起了广泛的关注，具有重要的分析与研究价值。

[0003] 深度卷积神经网络模型具有模型深度高、层次复杂、数据量级大、并行度高、计算密集和存储密集等特征，大批量的卷积运算和池化操作往往使其在应用过程当中成为一大计算瓶颈，大量中间结果的存储也对计算机存储结构提出了较高的要求，这对于实时性较强而投入成本有限的应用场景来说是十分不利的。

[0004] 当下比较常用的两种加速器是CPU和GPU，CPU基于其串行执行的结构特点在计算性能上并不能较理想地满足要求，GPU虽然在计算性能上优势明显但却与CPU一样无法突破功耗壁垒，并且CPU和GPU在可扩展性上都存在较为严重的限制。考虑到诸如上述因素，越来越多的人开始设计专用硬件系统来完成对深度卷积神经网络的加速，但如何结合硬件芯片特点和平台优势充分挖掘出深度卷积神经网络计算模型的并行性以及流水性，合理高效地充分利用有限硬件资源来完成设计仍是有待解决的问题。

## 发明内容

[0005] 本发明提供了一种适于硬件设计实现的深度卷积神经网络的计算方法，其目的在于同时结合深度卷积神经网络模型结构特点和硬件设计的特点及优势，对传统软件层中已有实现的卷积神经网络计算结构进行重新调整，充分挖掘其在计算过程当中潜在的并行性以及各计算层之间的流水性，使之更匹配于硬件设计的特点，以合理高效地充分利用有限资源，为深度卷积神经网络的硬件实现提供一种高效、可行且易于扩展的计算方法。

[0006] 本发明所提供的一种深度卷积神经网络的计算方法，其特征在于，该计算方法包括如下步骤：

[0007] STEP1：对于给定的深度卷积神经网络模型，在上层主机中利用深度卷积神经网络计算结构调整算法，预先生成与该深度卷积神经网络模型相对应的特定调整参数；

[0008] STEP2：将所述特定调整参数、训练好的卷积核参数从所述上层主机上加载置DDR片外存储器中；

[0009] STEP3：硬件设计系统从所述外存储器中直接加载全部的所述特定调整参数，完成所述特定调整参数的部署，利用所述特定调整参数调整深度卷积神经网络的计算结构；

[0010] STEP4：所述上层主机结合所述特定调整参数中提供的原始输入图像位置重排参数对输入图像进行像素点位置重排，并将重排后的图像加载至所述存储器中，接着启动深

度卷积神经网络的计算过程；

[0011] STEP5：所述硬件设计系统不断从所述DDR片外存储器中获取计算输入数据，在所述特定调整参数和所述卷积核参数的协同参与下完成相关计算过程。

[0012] 进一步地，所述特定调整参数分为两类：计算顺序序列化参数和填充过滤参数；所述计算顺序序列化参数为原始输入图像位置重排参数、新旧值选取标记参数和旧值选取地址参数；

[0013] 所述填充过滤参数为核参地址跳跃标记参数、跳跃目的核参地址参数和窗口计算提前结束标记参数；

[0014] 在深度卷积神经网络的池化层中，所述填充过滤参数单指窗口计算提前结束标记参数；

[0015] 其中，所述原始输入图像位置重排参数用于对所述上层主机中的输入图像进行像素点位置重排以得到重排后的图像；

[0016] 所述新旧值选取标记参数为其所在层的计算顺序序列化实现过程提供新旧值数据选取标记值指定，所述标记值指定是从上一层的特征图中顺序获取下一个参与计算的新值数据还是从已经获取过的新值数据中选取旧值数据，当所述新旧值选取标记参数指定从已经获取过的新值数据中选取旧值数据时，所述旧值选取地址参数为其提供选取旧值数据的地址；

[0017] 其中，所述核参地址跳跃标记参数在深度卷积神经网络的卷积层中指明当前计算位置之后是否存在填充元素，当存在填充元素时，则需要执行跳跃过滤操作，所述跳跃目的核参地址参数为其提供卷积核参数的跳跃目的地址，当一个原始计算窗口中存在填充元素时，由于跳跃过滤操作的存在，计算窗口中真正投入计算的元素数量将小于原始计算窗口大小，此时，所述窗口计算提前结束标记参数为其提供窗口计算提前结束标记。

[0018] 进一步地，所述STEP1中涉及的深度卷积神经网络计算结构调整算法，采用队列为主要数据结构遍历各层，以首层全连接层为起点，以起始输入图像层为终点，遍历过程中生成与每层相关的所述调整参数，每一层中的所有特征图在后续参与计算时共享与该层对应的一套所述特定调整参数；

[0019] 其中，所述结构调整算法的具体步骤如下：

[0020] STEP1-1以首层全连接层中所输入的单张特征图的元素排列顺序为初始排列顺序，并将表示该初始排列顺序的一维位置序号序列依次存入至队列中；

[0021] STEP1-2判断所述队列是否为空，为空时算法结束，否则转至下一步骤STEP1-3；

[0022] STEP1-3每次取队列队首位置序号进行扩充，根据所在层的神经元结构找到与该位置序号所在元素相对应的上层特征图中的计算窗口位置，并依次分析该计算窗口中的每个元素在其所在的单张特征图中的位置；

[0023] STEP1-4判断当前窗口是否分析完毕，若没有分析完毕，则转至步骤STEP1-5，否则，转至步骤STEP1-10；

[0024] STEP1-5分析下一个当前窗口中的元素，判断该元素是否处于所在特征图中的填充位置，若否，转至STEP1-6；否则转至STEP1-9；

[0025] STEP1-6为该层中的此次分析行为分配一个唯一的有效分析序号，所述有效分析序号从编号1开始依次递增分配，并判断其所对应位置的元素于其所在的单张特征图中是

否被首次分析到，若是，转至STEP1-7；否则转至STEP1-8；

[0026] STEP1-7将当前有效分析序号的新旧值选取标记值置为1，其中标记值为1表示选取新值；标记值为0表示选取旧值，并判断有效分析序号所对应位置的元素于是否处于起始输入图像层，若是，将当前有效分析序号添加到所述原始输入图像位置重排参数中；否则，将当前有效分析序号添加到所述队列队尾，转至所述STEP1-4；

[0027] STEP1-8将当前有效分析序号的新旧值选取标记值置为0，转至所述STEP1-4；

[0028] STEP1-9为该层中的此次分析行为分配一个唯一的无效分析序号，所述无效分析序号从编号1开始依次递增分配，并判断该无效分析序号是否位于一段连续无效分析序号的段首，若是，将其正前面的一个有效分析序号添加至所述核参地址跳跃标记参数中，将紧接在该段连续无效分析序号末尾的一个有效分析序号添加至所述跳跃目的核参地址参数中，转至所述STEP1-4；否则，直接转至所述STEP1-4；

[0029] STEP1-10判断分析完的计算窗口中是否出现过处于填充位置的元素，若是，将该计算窗口中最后一个有效分析序号添加至窗口计算提前结束标记参数中，转至所述STEP1-2；否则，直接转至所述STEP1-2。

[0030] 按照本方案实现的深度卷积神经网络计算结构调整算法，通过分析位于深度卷积神经网络中各个卷积层和池化层的神经元结构特性，根据后一层期望得到的单张特征图的元素排列顺序逆序推出前一层中对应参与计算的单张特征图的元素排列顺序，排列顺序以一维位置序号序列表示。该算法采用队列为主要数据结构遍历各层，以首层全连接层为起点，以起始输入图像层为终点，遍历过程中生成与每层相关的调整参数，每一层中的所有特征图在后续参与计算时共享与该层对应的一套调整参数。

[0031] 深度卷积神经网络计算结构调整算法，以首层全连接层中所输入的单张特征图的元素排列顺序为初始排列顺序，并将表示该初始排列顺序的一维位置序号序列依次存入至队列中，深度卷积神经网络计算结构调整算法每次取队列队首位置序号进行扩充，根据所在层的神经元结构找到与该位置序号所在元素相对应的上层特征图中的计算窗口位置，并依次分析该计算窗口中的每个元素在其所在的单张特征图中的位置，每一层中的每一次分析行为对应一个唯一的分析序号。当分析到的元素处于其所在的单张特征图中的填充位置时，该分析序号称为无效分析序号；否则，该分析序号称为有效分析序号。

[0032] 因而，每个无效分析序号都与上一层单张特征图中的一个填充位置的元素相对应，每个有效分析序号都与上一层单张特征图中的一个参与有效计算的非填充位置的元素相对应。

[0033] 每个有效分析序号都拥有与其相对应的新旧值选取标记，新旧值选取标记的取值有两个：选新值标记和选旧值标记。每个新旧值选取标记取值为选旧值标记的有效分析序号都额外拥有一个与之相对应的旧值选取地址，每个含有填充元素的计算窗口中的最后一个有效分析序号都额外拥有一个与之相对应的窗口计算提前结束标记。该层中所有新旧值选取标记的有序集合即为该层待求的新旧值选取标记参数；该层中所有旧值选取地址的有序集合即为该层待求的旧值选取地址参数；该层中所有窗口计算提前结束标记的有序集合即为该层待求的窗口计算提前结束标记参数。

[0034] 若该层为深度卷积神经网络中的卷积层，则该层中每一段连续的无效分析序号或单个成段的无效分析序号还需为其正前面的一个有效分析序号额外产生一个核参地址跳

跃标记和跳跃目的核参地址，跳跃目的核参地址即处于该段正后面的一个有效分析序号所对应位置的元素在其计算窗口中的位置序号。该层中所有核参地址跳跃标记的有序集合即为该层待求的核参地址跳跃标记参数；该层中所有跳跃目的核参地址的有序集合即为该层待求的跳跃目的核参地址参数。

[0035] 由于上一层不同计算窗口之间可能存在交集，因而不同的分析序号可能对应到上一层单张特征图中同一个位置的元素。

[0036] 当一个有效分析序号所对应位置的元素于其所在的单张特征图中被首次分析到时，则将此有效分析序号的新旧值选取标记取值为选新值标记，并将该元素在其所处的单张特征图中的一维位置序号添加到队列尾部，上一层所有被首次分析到的元素在其所处的单张特征图中的一维位置序号的有序集合即上一层单张特征图期望得到的元素排列顺序，根据求得的上一层单张特征图期望得到的元素排列顺序，按照上述方法，更进一步可以求得上上层单张特征图期望得到的元素排列顺序，直至求得起始图像输入层期望得到的元素排列顺序为止，起始图像输入层期望得到的元素排列顺序即待求的原始输入图像位置重排参数；

[0037] 当一个有效分析序号所对应位置的元素于其所在的单张特征图中并非被首次分析到时，则将此有效分析序号的新旧值选取标记取值为选旧值标记，并找到该元素在其所处的单张特征图中的一维位置序号在整张特征图期望得到的元素排列顺序中的位置，此位置即此有效分析序号额外拥有的旧值选取地址。

[0038] 与现有计算方式相比，本发明提供的方法更有利于深度卷积神经网络在专用硬件设计上的实现，按照本发明提供的方法能高效而充分地利用有限的硬件资源，低功耗、低成本地完成对深度卷积神经网络复杂计算模型的加速，在大幅度提高加速性能的同时还拥有灵活的可扩展性，能很好地满足以深度卷积神经网络的实现为基础且实时性要求较高的各类应用需求，在人工智能、机器学习、深度学习等领域有比较广泛的应用前景。主要的创新点如下：

[0039] (1)提出用相关调整参数对深度卷积神经网络的计算结构进行重新调整，打破了传统卷积神经网络中计算窗口结构固定的束缚，使得每一计算层先参与计算的数据能够先到达，充分挖掘出深度卷积神经网络中的计算并行性以及各计算层之间的流水性，有效地减少了大量中间结果的存储，使之更有利于在专用硬件设计上高效并行流水化实现。

[0040] (2)提出用相关调整参数自动过滤掉计算过程中存在的各类填充元素，在设计专用硬件系统完成深度卷积神经网络计算的过程中，能避免无效计算的投入，有效地解决深度卷积神经网络中由于各类填充操作而导致的资源浪费和有效计算延误问题。

[0041] (3)提出了一套生成所有相关调整参数的上层软件实现算法。

[0042] (4)提出了一整套调整后的深度卷积神经网络的高效并行流水化实现方案，包括内部各并行度的设置方法、存储优化策略等。

## 附图说明

[0043] 图1为本发明实现的硬件设计系统与上层主机之间的交互结构示意图；

[0044] 图2为本发明提出的深度卷积神经网络计算结构调整参数的结构框图；

[0045] 图3为本发明提出的深度卷积神经网络计算结构调整算法的数据处理流程图；

[0046] 图4为本发明实现的硬件设计系统的整体模块组成结构示意图；

[0047] 图5为按照本发明实现的硬件设计系统中卷积计算模块的数据处理示意图；

[0048] 图6为按照本发明实现的硬件设计系统中池化计算模块的数据处理示意图；

[0049] 图7为按照本发明实现的硬件设计系统中卷积计算顺序序列化实现模块的特征图元组选择功能子模块工作流程结构示意图；

[0050] 图8为按照本发明实现的硬件设计系统中卷积计算顺序序列化实现模块的卷积核参数选择功能子模块工作流程结构示意图；

[0051] 图9为按照本发明实现的硬件设计系统中的池化计算顺序序列化实现模块的组成结构示意图；

[0052] 图10为按照本发明实现的硬件设计系统中卷积计算模块的工作流程结构示意图；

[0053] 图11为按照本发明实现的硬件设计系统中的卷积核计算单元的实现原理图；

[0054] 图12为按照本发明实现的硬件设计系统中池化计算模块的工作流程结构示意图；

[0055] 图13为按照本发明实现的硬件设计系统中最大池化单元的实现原理图；

[0056] 图14为按照本发明实现硬件设计系统中的平均池化单元的实现原理图。

## 具体实施方式

[0057] 以下结合附图及实施例，对本发明进一步详细说明。应当理解，此处所描述的具体实施例仅仅用以解释本发明，并不用于限定本发明。

[0058] 作为具体实施例的深度卷积神经网络模型具有以下特点：

[0059] (1)所有计算层(计算层包括起始输入图像层、卷积层、池化层和全连接层)单张特征图的长宽相同，所有计算层计算窗口的长宽相同。

[0060] (2)各计算层的连接方式依次为：起始输入图像层、卷积层1、池化层1、卷积层2、池化层2、卷积层3、池化层3、全连接层1和全连接层2。

[0061] (3)池化操作仅存在两种方式：取最大值池化和取平均值池化；激活操作采用Relu激活方式。

[0062] (4)各计算层图像大小、图像填充大小、计算窗口大小、计算窗口移动步长和池化方式信息如下表所示：

[0063]

| 计算层名称 | 图像大小 | 填充大小 | 窗口大小 | 窗口步长 | 池化方式 |
|---|---|---|---|---|---|
| 输入图像层 | 32*32*3 | 2 | - | - | - |
| 卷积层1 | 32*32*32 | 0 | 5*5 | 1 | - |
| 池化层1 | 16*16*32 | 2 | 3*3 | 2 | 取最大值 |
| 卷积层2 | 16*16*32 | 0 | 5*5 | 1 | - |
| 池化层2 | 8*8*32 | 2 | 3*3 | 2 | 取平均值 |
| 卷积层3 | 8*8*64 | 0 | 5*5 | 1 | - |
| 池化层3 | 4*4*64 | 0 | 3*3 | 2 | 取平均值 |
| 全连接层1 | 1*1*64 | 0 | 1*1 | 0 | - |
| 全连接层2 | 1*1*10 | 0 | 1*1 | 0 | - |

[0064] (5)硬件设计系统上的存储资源能够存储任意连续两个卷积层用到所有卷积核参

数，但不能同时容纳所有卷积层的卷积核参数。

[0065] 如图1所示，整个深度卷积神经网络从模型参数的生成、部署到最终计算结果的回传，整个过程的处理流程如下：

[0066] A1.上层主机通过相关训练方法训练得到对应深度卷积神经网络模型的所有卷积核参数，这些卷积核参数将在后面将作为硬件设计系统中卷积操作实现部分的输入数据参与计算。

[0067] A2.上层主机调用本发明提出的深度卷积神经网络计算结构调整算法生成所有需要的调整参数。如图1中①、②所示。其中①表示将给定的深度卷积神经网络模型的模型参数作为输入数据送入到所述调整算法中，这些模型参数具体包括：深度卷积神经网络的计算层数信息、每一计算层单张特征图(起始输入图像层也看作是由多张特征图组成)的宽度信息、每一计算层计算窗口的宽度信息、每一计算层计算窗口移动步长信息、每一计算层特征图填充大小信息、每一计算层特征图元组大小(每一计算层所有参加计算的特征图在同一二维位置处的所有特征值的有序集合称为该二维位置处的特征图元组，特征图元组所包含的特征值个数称为特征图元组的大小)信息以及每一池化层的池化方式信息等。其中②表示通过所述调整算法生成所有相关调整参数。

[0068] A3.上层主机将生成的调整参数通过PCIe总线传送到板上的DDR片外存储器中，并在传送完毕后向硬件设计系统发送读调整参数命令，如图1中③所示；硬件设计系统接收到读调整参数命令后，启动DMA读操作通过PCIe总线从DDR片外存储器中获取调整参数并分别存入对应的硬件设计系统存储器中。

[0069] A4.将所述训练好的卷积核参数通过PCIe总线送入板上的DDR片外存储器中，并在传送完毕后向硬件设计系统发送读卷积核参数命令，如图1中④所示。由于硬件设计系统上的存储资源不能一次性容纳所有的卷积核参数，在接收到读调整参数命令后，硬件设计系统启动DMA读操作通过PCIe总线从DDR片外存储器中预先获取前两个卷积层所用到的卷积核参数存入硬件设计系统上的卷积核参数存储器中，其它卷积层所用到的卷积核参数将在计算过程中适时地分批次加载。

[0070] A5.上层主机通过生成的所述调整参数中的原始输入图像位置重排参数对所有输入图像进行像素点位置重排，如图1中⑤所示；并将重排后的图像通过PCIe总线送入板上的DDR片外存储器中，传送完毕后向硬件设计系统发送计算启动命令，如图1中⑥所示。

[0071] A6.硬件设计系统在收到计算启动命令后，启动DMA读操作通过PCIe总线从DDR片外存储器中获取重排后的图像数据开始计算，计算过程中，硬件设计系统需要多次适时地从DDR片外存储器继续获取其他卷积层的卷积核参数，在调整参数和卷积核参数的协同参与下完成相关计算过程。待到生成相关计算结果后，再启动DMA写操作将计算结果回传到DDR片外存储器中，并向上层主机发送计算完成中断通知，如图1中⑦所示。

[0072] A7.上层主机接收到硬件设计系统发送的计算完成中断通知后，从DDR片外存储器的指定位置读取计算结果继而进行后续所需操作，如图1中⑧所示。

[0073] 如图2所示，调整参数主要分为两类：计算顺序序列化参数和填充过滤参数。其中，计算顺序序列化参数可进一步细分为原始输入图像位置重排参数、新旧值选取标记参数和旧值选取地址参数；在深度卷积神经网络的卷积层中，填充过滤参数可进一步细分为核参地址跳跃标记参数、跳跃目的核参地址参数和窗口计算提前结束标记参数；在深度卷积神

经网络的池化层中，填充过滤参数单指窗口计算提前结束标记参数。

[0074] 计算顺序序列化参数打破了传统卷积神经网络中计算窗口结构固定的束缚，使得每一计算层先参与计算的数据能够先到达，充分挖掘出深度卷积神经网络中的计算并行性以及层与层之间的流水性，有效地减少了大量中间结果的存储，使之更有利于在专用硬件设计上高效并行流水化实现。其中，原始输入图像位置重排参数用于对上层主机中的输入图像进行像素点位置重排以得到重排后的图像；新旧值选取标记参数为其所在层的计算顺序序列化实现过程提供新旧值数据选取标记，标记值指定是从上一层的特征图(起始输入图像层也看作是由多张特征图组成)中顺序获取下一个参与计算的新值数据还是从已经获取过的新值数据中选取旧值数据。当新旧值选取标记参数指定从已经获取过的新值数据中选取旧值数据时，旧值选取地址参数为其提供选取旧值数据的地址。

[0075] 填充过滤参数针对深度卷积神经网络的卷积层中可能存在的特征图尺寸填充现象和池化层中可能存在的窗口越界填充现象所带来的无效计算问题，在设计专用硬件系统实现计算的过程中，能自动过滤掉填充元素，避免无效计算的投入，有效地解决深度卷积神经网络中由于各类填充操作而导致的资源浪费和有效计算延误问题。其中核参地址跳跃标记参数在深度卷积神经网络的卷积层中指明当前计算位置之后是否存在填充元素，当存在填充元素时，则需要执行跳跃过滤操作，跳跃目的核参地址参数为其提供卷积核参数的跳跃目的地址。当一个原始计算窗口中存在填充元素时，由于跳跃过滤操作的存在，计算窗口中真正投入进计算的元素数量将小于原始计算窗口大小，此时，窗口计算提前结束标记参数为其提供窗口计算提前结束标记。

[0076] 深度卷积神经网络计算结构调整算法，通过分析位于深度卷积神经网络中各个卷积层和池化层的神经元结构特性，根据后一层期望得到的单张特征图的元素排列顺序逆序推出前一层中对应参与计算的单张特征图的元素排列顺序，排列顺序以一维位置序号序列表示。该算法采用队列(记为Q)为主要数据结构遍历各层，以首层全连接层为起点，以起始输入图像层为终点，遍历过程中生成与每层相关的调整参数，每一层中的所有特征图在后续参与计算时共享与该层对应的一套调整参数。

[0077] 深度卷积神经网络计算结构调整算法，以首层全连接层中所输入的单张特征图的元素排列顺序为初始排列顺序，并将表示该初始排列顺序的一维位置序号序列依次存入至所述队列中，深度卷积神经网络计算结构调整算法每次取队列队首位置序号进行扩充，根据所在层的神经元结构找到与该位置序号所在元素相对应的上层特征图中的计算窗口位置，并依次分析该计算窗口中的每个元素在其所在的单张特征图中的位置，每一层中的每一次分析行为对应一个唯一的分析序号。当分析到的元素处于其所在的单张特征图中的填充位置时，该分析序号称为无效分析序号；否则，该分析序号称为有效分析序号。

[0078] 因而，每个无效分析序号都与上一层单张特征图中的一个填充位置的元素相对应，每个有效分析序号都与上一层单张特征图中的一个参与有效计算的非填充位置的元素相对应。

[0079] 每个有效分析序号都拥有与其相对应的新旧值选取标记，新旧值选取标记的取值有两个：选新值标记和选旧值标记。每个新旧值选取标记取值为选旧值标记的有效分析序号都额外拥有一个与之相对应的旧值选取地址，每个含有填充元素的计算窗口中的最后一个有效分析序号都额外拥有一个与之相对应的窗口计算提前结束标记。该层中所有新旧值

选取标记的有序集合即为该层待求的所述新旧值选取标记参数；该层中所有旧值选取地址的有序集合即为该层待求的所述旧值选取地址参数；该层中所有窗口计算提前结束标记的有序集合即为该层待求的所述窗口计算提前结束标记参数。

[0080] 若该层为深度卷积神经网络中的卷积层，则该层中每一段连续的无效分析序号或单个成段的无效分析序号还需为其正前面的一个有效分析序号额外产生一个核参地址跳跃标记和跳跃目的核参地址，跳跃目的核参地址即处于该段正后面的一个有效分析序号所对应位置的元素在其计算窗口中的位置序号。该层中所有核参地址跳跃标记的有序集合即为该层待求的所述核参地址跳跃标记参数；该层中所有跳跃目的核参地址的有序集合即为该层待求的所述跳跃目的核参地址参数。

[0081] 由于上一层不同计算窗口之间可能存在交集，因而不同的分析序号可能对应到上一层单张特征图中同一个位置的元素。

[0082] 当一个有效分析序号所对应位置的元素于其所在的单张特征图中被首次分析到时，则将此有效分析序号的新旧值选取标记取值为选新值标记，并将该元素在其所处的单张特征图中的一维位置序号添加到所述队列尾部，上一层所有被首次分析到的元素在其所处的单张特征图中的一维位置序号的有序集合即上一层单张特征图期望得到的元素排列顺序，根据求得的上一层单张特征图期望得到的元素排列顺序，按照上述方法，更进一步可以求得上上层单张特征图期望得到的元素排列顺序，直至求得起始图像输入层期望得到的元素排列顺序为止，起始图像输入层期望得到的元素排列顺序即待求的所述原始输入图像位置重排参数；

[0083] 当一个有效分析序号所对应位置的元素于其所在的单张特征图中并非被首次分析到时，则将此有效分析序号的新旧值选取标记取值为选旧值标记，并找到该元素在其所处的单张特征图中的一维位置序号在整张特征图期望得到的元素排列顺序中的位置，此位置即此有效分析序号额外拥有的旧值选取地址。

[0084] 如图3所示，算法的数据处理流程如下：

[0085] A1.以首层全连接层中所输入的单张特征图的元素排列顺序为初始排列顺序，并将表示该初始排列顺序的一维位置序号序列依次存入队列Q中。此实施例中首层全连接层中所输入的单张特征图大小对应到前一层池化层3生成的特征图二维大小为4*4，由于全连接层只有一个计算窗口，所以输入的单张特征图的元素排列顺序为1～16；因而将1～16依次存入Q中。

[0086] A2.判断队列Q是否为空，为空时，算法结束；否则，转至A3；

[0087] A3.取队列Q队首位置序号进行扩充，根据所在层的神经元结构找到与该位置序号所在元素相对应的上层特征图中的计算窗口位置，并依次分析该计算窗口中的每个元素在其所在的单张特征图中的位置。例如第一次取出的列队首位置序号为1，对应到卷积层3生成的特征图中大小为3*3，步长为1的1号计算窗口，因而接下来将依次分析1号计算窗口中的元素，具体对应到卷积层3中生成的单张特征图中一维位置序号为1、2、3、9、10、11、17、18、19的元素。

[0088] A4.判断当前窗口是否分析完毕，若没有分析完毕，转至A5；否则，转至A10；

[0089] A5.分析下一个当前窗口中的元素，判断该元素是否处于所在特征图中的填充位置。若否，转至A6；否则转至A9。

[0090] A6.为该层中的此次分析行为分配一个唯一的有效分析序号，有效分析序号从编号1开始依次递增分配，并判断该有效分析序号所对应位置的元素于其所在的单张特征图中是否被首次分析到，若是，转至A7；否则转至A8。

[0091] A7.将当前有效分析序号的新旧值选取标记值置为1(标记值为1表示选取新值；标记值为0表示选取旧值)。并判断有效分析序号所对应位置的元素于是否处于起始输入图像层，若是，将当前有效分析序号添加到原始输入图像位置重排参数中；否则，将当前有效分析序号添加到队列Q队尾。转至A4。

[0092] A8.将当前有效分析序号的新旧值选取标记值置为0，转至A4。

[0093] A9.为该层中的此次分析行为分配一个唯一的无效分析序号，无效分析序号从编号1开始依次递增分配，并判断该无效分析序号是否位于一段连续无效分析序号的段首，若是，将其正前面的一个有效分析序号添加至核参地址跳跃标记参数中，将紧接在该段连续无效分析序号末尾的一个有效分析序号添加至跳跃目的核参地址参数中，转至A4；否则，直接转至A4。

[0094] A10.判断分析完的计算窗口中是否出现过处于填充位置的元素，若是，将该计算窗口中最后一个有效分析序号添加至窗口计算提前结束标记参数中，转至A2。否则，直接转至A2。

[0095] 如图4所示，按照本发明中的深度卷积神经网络的计算方法实现的硬件设计系统主要由输入数据分配控制模块、输出数据分配控制模块、卷积计算顺序序列化实现模块、池化计算顺序序列化实现模块、卷积计算模块、池化计算模块和卷积计算结果分配控制模块七大模块组成，此外硬件设计系统还包含一个内部系统级联接口。

[0096] 输入数据分配控制模块同时与硬件设计系统外围接口和所述内部系统级联接口、卷积计算顺序序列化实现模块相连；输出数据分配控制模块同时与硬件设计系统外围接口和所述内部系统级联接口、卷积计算结果分配控制模块以及池化计算模块相连；卷积计算结果分配控制模块同时与卷积计算模块、输出数据分配控制模块以及池化计算顺序序列化实现模块相连；卷积计算顺序序列化实现模块与卷积计算模块之间直接相连；池化计算顺序序列化实现模块与池化计算模块之间直接相连。

[0097] 输入数据分配控制模块主要负责实时监视卷积计算顺序序列化实现模块的数据消耗状况，适时适量地向DDR片外存储器发送相关读数据命令并及时接收硬件设计系统外围接口和所述内部系统级联接口传送来的输入数据，除此之外，输入数据分配控制模块还需将接收到的数据有组织有规格地传送给卷积计算顺序序列化实现模块。

[0098] 输出数据分配控制模块主要负责及时接收池化计算模块或卷积计算结果分配控制模块传送来的输入数据，并根据当前所处的计算阶段将接收到的数据有组织有规格地传送给所述内部系统级联接口或硬件设计系统外围接口，适时适量地向DDR片外存储器发送相关写数据命令和相关中断通知。除此之外，输出数据分配控制模块还负责实时响应硬件设计系统外围接口传送来的各类相关命令。

[0099] 卷积计算顺序序列化实现模块主要负责结合相关调整参数将深度卷积神经网络中相关卷积操作的结构化计算顺序序列化，并为卷积计算模块及时传送序列化后的数据集；池化计算顺序序列化实现模块主要负责结合相关调整参数将深度卷积神经网络中相关池化操作的结构化计算顺序序列化，并为池化计算模块及时传送序列化后的数据集。

[0100] 卷积计算模块主要负责完成深度卷积神经网络中的相关卷积计算，并将计算结果及时传送给卷积计算结果分配控制模块；池化计算模块主要负责完成深度卷积神经网络中的相关池化操作，并将计算结果及时传送给所述输出数据分配控制模块。

[0101] 卷积计算结果分配控制模块主要负责及时接收卷积计算模块传来的计算结果数据，并根据当前所处的计算阶段将接收到的数据有组织有规格地传送给池化计算顺序序列化实现模块或输出数据分配控制模块。

[0102] 内部系统级联接口主要负责为硬件设计系统内部子系统之间的级联或内部模块之间的连接提供有效接口，用于连接输出数据分配控制模块和输入数据分配控制模块。

[0103] 在硬件设计系统中的各层计算过程中，每一计算层所有参加计算的特征图在同一二维位置处的所有特征值的有序集合称为该二维位置处的特征图元组，特征图元组所包含的特征值个数称为特征图元组的大小。特征图元组将作为一个整体先后参与计算，原始输入图像层的处理也按照将其视为特征图的方式进行，二维位置计算点的移动由上一计算层或起始图像输入层的数据送出顺序和卷积计算顺序序列化实现模块或池化计算顺序序列化实现模块联合决定；每一计算层中所有生成的特征图也以特征图元组为基本单位依次生成，上一个特征图元组生成完毕后才开始进行下一个特征图元组的生成。输入的特征图元组大小记作DIN，生成的特征图元组大小记作DON。

[0104] 上层主机根据所述调整参数中提供的原始输入图像位置重排参数对输入图像进行像素点位置重排，无论是在重排过程中还是在之后的重排图像的数据传送过程中，图像的各三维分量皆作为一个整体进行操作。重排后的图像根据图像二维大小，按照从左到右，从上到下的顺序依次传送给DDR片外存储器。上层主机中的卷积核参数按照所述卷积计算模块设定的计算顺序按规格组织后再传送给DDR片外存储器。

[0105] 输入数据分配控制模块、输出数据分配控制模块和卷积计算结果分配控制模块在传送数据时皆保持其数据接收的先后顺序不变，仅当接收到的数据组成一定大小的数据单元后即将其发送给与其相连的所需模块。

[0106] 卷积计算模块每次同时并行处理多张特征图，每张特征图每次同时并行与多个卷积核进行卷积操作，因而卷积计算模块可以每次同时并行生成多张新的特征图；池化计算模块同样每次同时并行处理多张特征图。卷积计算模块每次最多同时处理的特征图张数称为卷积层特征图并行度，记作KFP；卷积计算模块每次最多同时生成的特征图张数称为卷积核组并行度，记作KGP；池化计算模块每次最多同时处理的特征图张数称为池化层特征图并行度，记作PFP。

[0107] 卷积计算模块的数据处理示意图如图5所示，其中if1～ifn代表上层生成输入的n张特征图，of1～ofn代表本层生成的n张特征图；其中连接输入特征图与卷积核参数阵列的⊗符号表示乘法操作，连接各⊗符号与生成特征图元素的⊕符号代表加法操作。在深度卷积神经网络的全连接层中，图中输入的特征图和生成的特征图只包含一个特征图元素，计算窗口大小将等于整张输入特征图的大小。

[0108] 池化计算模块的数据处理示意图如图6所示，其中if1～ifn代表上层生成输入的n张特征图，of1～ofn代表本层生成的n张特征图；其中连接输入特征图的计算窗口与生成特征图元素的Ⓟ符号代表池化操作。

[0109] 特征图元组的每次选择操作与有效分析序号一一对应。

[0110] ①特征图元组选择功能子模块

[0111] 如图7所示，特征图元组选择功能子模块主要由特征图元组存储器、新旧选择器、标记参数存储器、地址参数存储器、计算窗口缓冲存储器和特征图元组计数器组成。

[0112] 其中，特征图元组存储器采用双端口RAM实现，用于存储所述输入数据分配控制模块送入的特征图元组；新旧选择器维护两个地址寄存器，分别为新值地址寄存器和旧值地址寄存器，用于从特征图元组存储器中选择相应的特征图元组并输出给所述卷积计算模块；标记参数存储器用于存储所述的有效分析序号的新旧值选取标记和窗口计算提前结束标记，地址参数存储器用于存储所述的有效分析序号的旧值选取地址，对于一个给定的深度卷积神经网络模型，标记参数存储器和地址参数存储器一次写入多次循环读取；计算窗口缓冲存储器采用双端口RAM实现，用于缓存新旧选择器输出的特征图元组并将其输出给所述卷积计算模块；特征值图元组计数器用于统计新旧选择器选择输出的特征图元组个数。

[0113] 特征图元组选择功能子模块每节拍从所述输入数据分配控制模块获取一个特征图元组的KFP个特征值，这KFP个特征值组成一个输入特征值组。新旧选择器每次选择特征图元组进行输出时，查看当前新旧值选取标记值，若当前新旧值选取标记值为选新值标记，则从新值地址寄存器提供的起始地址处开始以特征值组为单位进行特征图元组的输出，每输出一个特征组后，新值地址寄存器自动加一，当当前选取的特征图元组输出完毕后，从标记参数存储器中顺序获取下一个新旧值选取标记作为当前新旧值选取标记；若当前新旧值选取标记值为选旧值标记，则将当前旧值选取地址送入旧值地址寄存器，并以此地址为起始地址以特征值组为单位进行特征图元组的输出，每输出一个特征组后，旧值地址寄存器自动加一，当当前选取的特征图元组输出完毕后，从标记参数存储器中顺序获取下一个新旧值选取标记作为当前新旧值选取标记，并从地址参数存储器中顺序获取下一个旧值选取地址作为当前旧值选取地址。每当新旧选择器输出完一个特征图元组后，特征图元组计数器自动加一，若此时新旧选择器选择输出的特征图元组达到一个无填充元素的计算窗口大小，新旧选择器将暂停输出，直至位于计算窗口缓冲存储器中的当前计算窗口的特征图元组重复使用((DON-1)/KGP+1)次为止；若此时前新旧选择器选择输出的特征图元组尚未达到一个无填充元素的计算窗口大小，但当前特征图元组计数器值与当前窗口计算提前结束标记值相同，此时新旧选择器也将提前暂停输出，直至位于计算窗口缓冲存储器中的当前计算窗口的特征图元组重复使用((DON-1)/KGP+1)次为止，并且在新旧选择器提前暂停输出的同时，从标记参数存储器中顺序获取下一个窗口计算提前结束标记作为当前窗口计算提前结束标记。

[0114] ②卷积核参数选择功能子模块

[0115] 卷积核参数选择功能子模块中卷积核参数阵列的输出与所述特征图元组选择功能子模块中输出特征值组的输出同步进行。

[0116] 如图8所示，卷积核参数选择功能子模块主要由卷积核参数存储器(a)、卷积核参数存储器(b)、选择器、标记参数存储器、地址参数存储器和核参阵列组计数器组成。

[0117] 其中，卷积核参数存储器(a)和卷积核参数存储器(b)采用双端口RAM实现，用于存储所述输入数据分配控制模块送入的卷积核参数；标记参数存储器用于存储所述的核参地址跳跃标记参数，地址参数存储器用于存储所述的跳跃目的核参地址参数，对于一个给定

的深度卷积神经网络模型，标记参数存储器和地址参数存储器一次写入多次循环读取；选择器维护一个地址寄存器和一个跳转地址生成器，用于从卷积核参数存储器(a)或卷积核参数存储器(b)中选择相应的卷积核参数阵列组(与所述特征图元组选择功能子模块中输出的一个特征图元组相对应的所有卷积核参数阵列的集合称为一个卷积核参数阵列组)输出给所述卷积计算模块，其中跳转地址生成器从地址参数存储器获取跳跃目的核参地址参数进行计算，为选择器提供对应的跳跃目的核参地址；核参阵列组计数器用于统计输出的卷积核参数阵列组个数。

[0118] 选择器每次选择卷积核参数阵列组进行输出时，比较当前核参地址跳跃标记参数值与当前核参阵列组计数器计数值是否相等。若相等，则将所述跳转地址生成器的当前跳转地址送入地址寄存器，并以此地址为起始地址，以卷积核参数阵列为单位进行卷积核参数阵列组的输出，每输出一个卷积核参数阵列，地址寄存器自动加一，当当前选取的卷积核参数阵列组输出完毕后，核参阵列组计数器自动增一，所述跳转地址生成器计算输出下一个跳转地址作为当前跳转地址；若不相等，则直接从所述地址寄存器提供的起始地址处开始，以卷积核参数阵列为单位进行卷积核参数阵列组的输出，每输出一个卷积核参数阵列，地址寄存器自动加一，当当前选取的卷积核参数阵列组输出完毕后，核参阵列组计数器自动增一。在选择器选择卷积核参数阵列组进行输出的过程中，卷积核参数存储器(a)和卷积核参数存储器(b)轮流切换为选择器提供卷积参数阵列组，切换操作发生当前计算层结束时刻，从所述输入数据分配控制模块送入的卷积核参数也以计算层为单位轮流依次送入卷积核参数存储器(a)和卷积核参数存储器(b)。

[0119] 池化计算顺序序列化实现模块获取特征图元组的操作与卷积计算顺序序列化实现模块的获取过程类似，但每节拍获取的特征图元组的特征值个数为PFP，并且当当前窗口计算结束时，计算窗口中的所有特征图元组不需要重复参与计算。

[0120] 如图9所示，池化计算顺序序列化实现模块主要由特征图元组存储器、新旧选择器、标记参数存储器、地址参数存储器和特征图元组计数器组成。

[0121] 其中，特征图元组存储器采用双端口RAM实现，用于存储所述输入数据分配控制模块送入的特征图元组；新旧选择器维护两个地址寄存器，分别为新值地址寄存器和旧值地址寄存器，用于从特征图元组存储器中选择相应的特征图元组并输出给所述卷积计算模块；标记参数存储器用于存储所述的有效分析序号的新旧值选取标记和窗口计算提前结束标记，地址参数存储器用于存储所述的有效分析序号的旧值选取地址，对于一个给定的深度卷积神经网络模型，标记参数存储器和地址参数存储器一次写入多次循环读取；特征值图元组计数器用于统计新旧选择器选择输出的特征图元组个数。

[0122] 池化计算顺序序列化实现模块每节拍从所述输入数据分配控制模块获取一个特征图元组的PFP个特征值，这PFP个特征值组成一个输入特征值组。新旧选择器每次选择特征图元组进行输出时，查看当前新旧值选取标记值，若当前新旧值选取标记值为选新值标记，则从新值地址寄存器提供的起始地址处开始以特征值组为单位进行特征图元组的输出，每输出一个特征组后，新值地址寄存器自动加一，当当前选取的特征图元组输出完毕后，从标记参数存储器中顺序获取下一个新旧值选取标记作为当前新旧值选取标记；若当前新旧值选取标记值为选旧值标记，则将当前旧值选取地址送入旧值地址寄存器，并以此地址为起始地址以特征值组为单位进行特征图元组的输出，每输出一个特征组后，旧值地

址寄存器自动加一，当当前选取的特征图元组输出完毕后，从标记参数存储器中顺序获取下一个新旧值选取标记作为当前新旧值选取标记，并从地址参数存储器中顺序获取下一个旧值选取地址作为当前旧值选取地址。每当新旧选择器输出完一个特征图元组后，特征图元组计数器自动加一，若此时新旧选择器选择输出的特征图元组未达到一个无填充元素的计算窗口大小，但当前特征图元组计数器值与当前窗口计算提前结束标记值相同，此时所述池化计算顺序序列化实现模块向所述池化计算模块发送当前窗口计算提前结束信号，并从标记参数存储器中顺序获取下一个窗口计算提前结束标记作为当前窗口计算提前结束标记。

[0123] 所述卷积计算顺序序列化实现模块和所述池化计算顺序序列化实现模块中的所述特征图元组存储器在其所在计算层中进行分时循环利用，所述特征图元组存储器并不为上一层传送过来的每一特征图元组都单独提供存储单元，其容量大小的设定结合所在计算域中同一特征图元组新值存入和旧值重取之间的最大地址间隔给出；

[0124] 旧值选取地址参数在经所述上层主机传送到所述DDR片外存储器之前需做相应的取余操作，取余模长为其所在计算域的所述特征图元组存储器容量大小。

[0125] 如图10所示，卷积计算模块由KGP(图中m＝KGP)个卷积核计算单元并列组成。

[0126] 卷积计算模块在每一个有效节拍同时获取卷积计算顺序序列化实现模块传入的KFP个特征值与KFP*KGP个卷积核参数，这些卷积核参数来自KGP个不同的卷积核。获取到的KFP个特征值将同时与这KGP个卷积核进行卷积操作，卷积计算结果加上相应的偏置值再经过Relu激活操作后，得到KGP个特征图元素，这KGP个特征图元素对应属于KGP张不同的生成特征图并且最终会被依次送往卷积计算结果分配控制模块。

[0127] 如图11所示，卷积核计算单元主要由乘加树、加法树、加偏器和激活器组成。乘加树由若干乘法器和加法器互连组成，加法树由若干加法器互连组成。

[0128] 其中乘加树、加法树共同完成卷积计算单元中的乘累加操作，加偏器完成卷积计算单元中的偏置相加操作，激活器完成卷积计算单元中的激活操作。

[0129] 卷积核计算单元在每个有效节拍同时获取来自所述卷积核参数选择功能子模块的KFP个特征值和来自所述卷积核参数选择功能子模块的KFP个卷积核参数。乘加树对KFP个特征值和KFP卷积核参数进行乘累加操作，并将乘累加结果按序依次送入加法树中进行二次集中累加。待到加法树首层入口处的操作数全部就绪或当前计算窗口的最后一组特征值就绪后，加法树启动计算完成二次累加；待到当前计算窗口的全部累加操作完成，加法树将最后的累加结果送入加法器中进行偏置相加操作，偏置相加操作完成后，相加结果继而会被送入激活器进行激活，激活后的结果即卷积计算单元的最终计算结果。卷积计算单元的最终计算结果将被送入所述卷积计算结果分配控制模块。

[0130] 卷积计算单元中的加法树主要用于缓存乘加树送入的乘累加结果，并集中进行累加计算，加法树的二次集中累加有效地解决了在浮点数累加过程中，由于前后操作数的数据相关性而引发的流水线断流，继而导致的卷积核计算单元取数阻塞问题，有效地缓解了深度卷积神经网络中处于卷积计算部分的一大计算瓶颈障碍。

[0131] 如图12所示，池化计算模块主要由分配器、最大值池化单元、平均值池化单元和选择器组成；

[0132] 池化计算模块在每个有效节拍同时获取来自所述池化计算顺序序列化实现模块

的PFP个特征值，并将该输入特征值组送入分配器进行分配；分配器则根据当前计算层的池化方式将输入的特征图元组分配给最大值池化单元或平均值池化单元；其中，最大值池化单元取每张特征图中当前计算窗口的最大特征图元素进行池化，平均值池化单元取每张特征图中当前计算窗口的所有特征图元素平均值进行池化；池化操作完成后，选择器根据当前计算层的池化方式选择最大值池化单元或平均值池化单元的池化结果送给所述输出数据分配控制模块。

[0133] 如图13所示，最大值池化单元主要由比较器阵列、中间结果缓存队列、分配器和特征图元组计数器组成。比较器阵列由若干比较器组成。

[0134] 其中，比较器阵列用于完成比较每张特征图中当前计算窗口的所有特征值元素，求取其最大值；中间结果缓存队列用于缓存比较器阵列比较的中间结果；分配器用于分配中间结果缓存队列中的中间结果，根据相关控制条件，将其送入比较器阵列进行迭代比较或将其作为最终结果输出给所述池化计算模块中的选择器；特征图元组计数器用于统计送入比较器阵列参与比较计算的特征图元组个数。

[0135] 最大值池化单元在每个有效节拍同时获取来自所述池化计算模块分配器的PFP个特征值，并将该输入特征值组送入比较器阵列，当一个特征图元组送入完毕后，特征图元组计数器自动加一；与此同时，分配器从中间结果缓存队列获取与输入特征值相对应的中间结果特征值组送入比较器阵列。一旦比较器阵列操作数准备就绪，比较器阵列启动计算，比较两组特征值组中各特征值分量，取其较大者送入中间结果缓存队列。当特征图元组计数器数值达到当前计算窗口大小时，分配器将位于中间结果缓存队列中的结果作为输出送入所述池化计算模块中的选择器。

[0136] 如图14所示，平均值池化单元主要由加法器阵列、中间结果缓存队列、分配器、特征图元组计数器和除法器阵列组成。加法器阵列由若干加法器组成，除法器阵列由若干除法器组成。

[0137] 其中，加法器阵列用于完成累加输入的特征图元组；中间结果缓存队列用于缓存加法器阵列累加的中间结果；分配器用于分配中间结果缓存队列中的中间结果，根据相关控制条件，将其送入加法器阵列进行迭代累加或将其作为最终结果输出给所述池化计算模块中的选择器；特征图元组计数器用于统计送入加法器阵列参与比较计算的特征图元组个数；除法器用于对分配器送出的累加结果进行取平均值操作。

[0138] 平均值池化单元在每个有效节拍同时获取来自所述池化计算模块分配器的PFP个特征值，并将该输入特征值组送入加法器阵列，当一个特征图元组送入完毕后，特征图元组计数器自动加一；与此同时，分配器从中间结果缓存队列获取与输入特征值相对应的中间结果特征值组送入加法器阵列。一旦加法器阵列操作数准备就绪，加法器阵列启动计算，完成两组特征值组中各特征值分量的累加，累加结果送入中间结果缓存队列。当特征图元组计数器数值达到当前计算窗口大小时，分配器将位于中间结果缓存队列中的结果送入除法器阵列；与此同时特征图元组计数器的当前数值也送入除法器阵列作为操作数参与计算，除法器阵列输出的平均值将作为输出送入池化计算模块中的选择器。

[0139] KFP、KGP的设定值结合给定的深度卷积神经网络模型中各卷积层的DON和硬件设计时的各类可用资源数量联合给出，在各类可用资源数量允许的情况下，尽量将KFP、KGP向所有卷积层中最大的DON靠近；PFP的设定值在保证紧接其后的卷积层不空闲的前提下尽量

减小。在本实施例中，KFP、KGP值均设定为8，PFP值设定为1。

[0140] 当KFP的值增大到一定程度之后，若此相关可用硬件资源依旧充足，则可利用内部系统级联接口对已有硬件设计系统进行进一步扩展。扩展后的硬件设计系统由多个硬件设计子系统级联而成，而每个硬件设计子系统皆由所述的七大模块和一个内部系统级联接口组成，其中，内部系统级联接口用于连接上一个硬件设计子系统的输出数据分配控制模块和下一个硬件设计子系统的输入数据分配控制模块，而七大模块之间的连接及实现除所在计算域和分析域有所缩减以外，与扩展前的硬件设计系统完全相同。

[0141] 扩展后的硬件设计系统不仅能成倍地提高计算并行度，合理地利用剩余硬件资源，而且能更充分地利用深度卷积神经网络中计算层层与层之间的流水性，有效缩短池化层与卷积层之间由于卷积层的计算瓶颈而带来的非必要等待时间，非必要等待时间的缩短意味着非必要中间结果的进一步减少，硬件设计时的可用存储资源将得到更为高效而充分的利用。

[0142] 本领域的技术人员容易理解，以上所述仅为本发明的较佳实施例而已，并不用以限制本发明，凡在本发明的精神和原则之内所作的任何修改、等同替换和改进等，均应包含在本发明的保护范围之内。

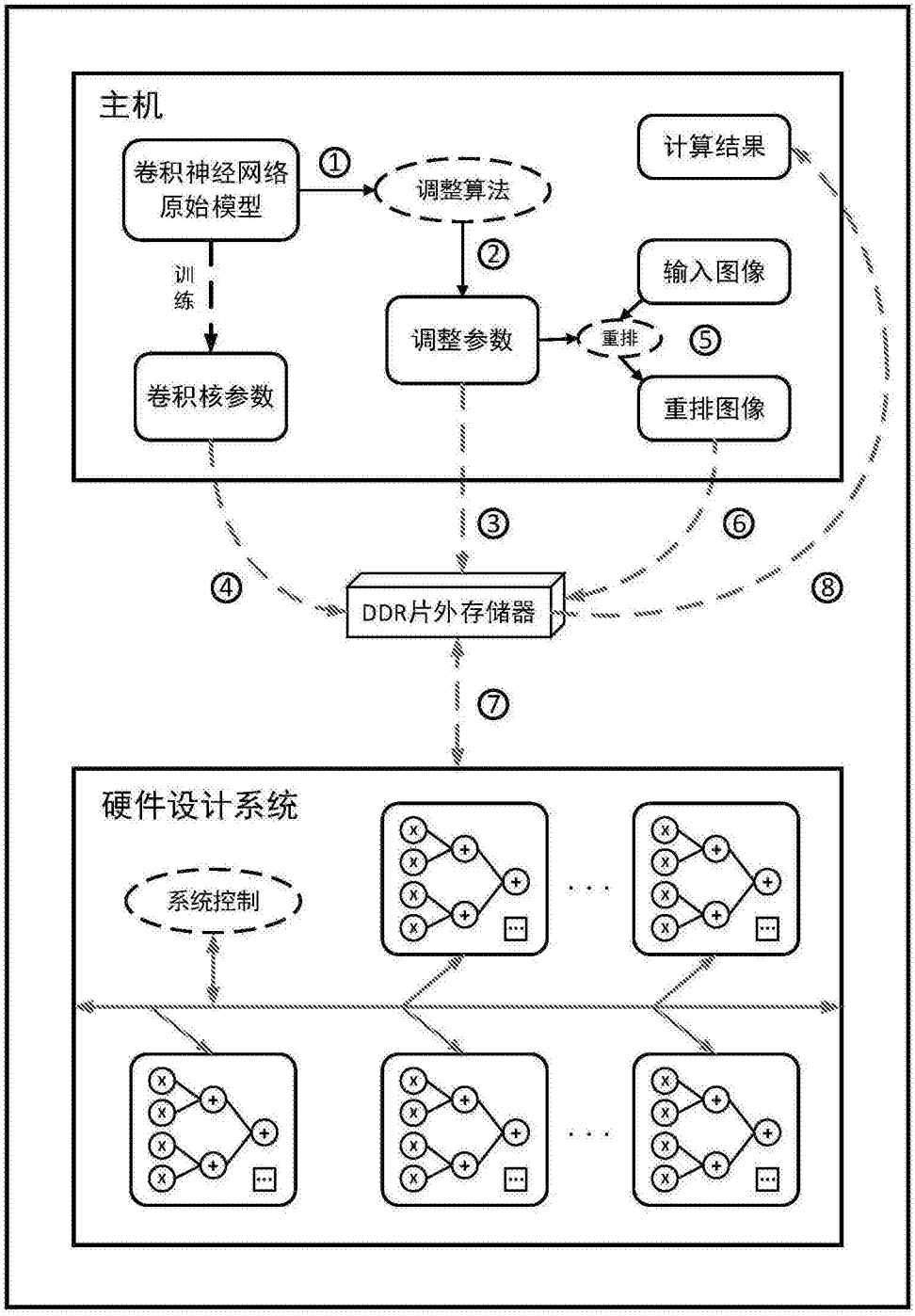

主机
卷积神经网络
原始模型
①
调整算法
计算结果
②
输入图像
训
练
调整参数
重排
⑤
卷积核参数
重排图像
③
⑥
④
DDR片外存储器
⑧
⑦
硬件设计系统
系统控制

图1

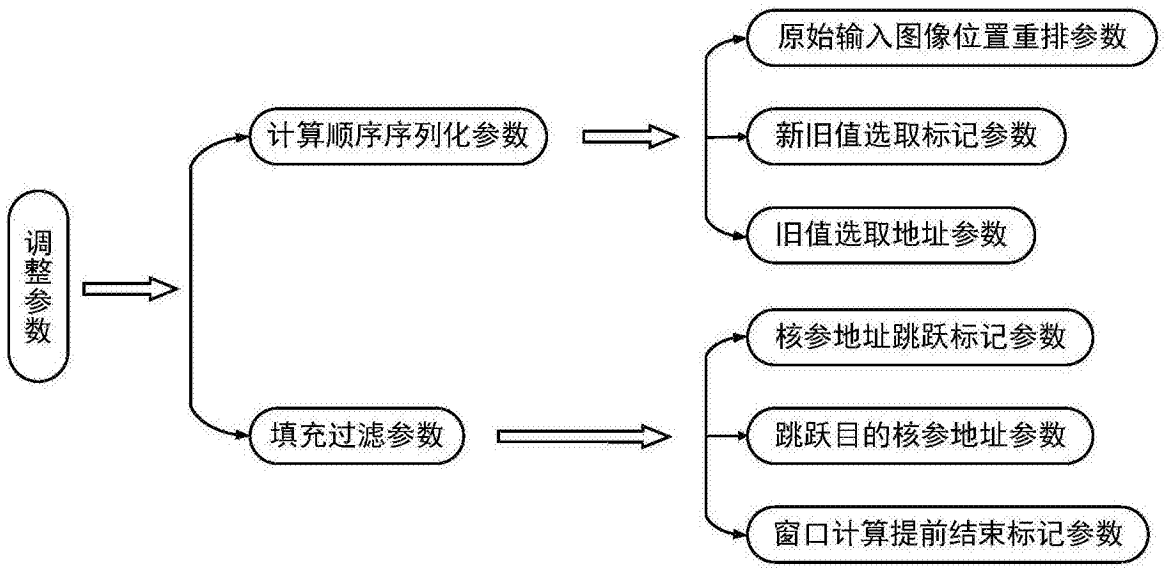

调整参数
计算顺序序列化参数
填充过滤参数
原始输入图像位置重排参数
新旧值选取标记参数
旧值选取地址参数
核参地址跳跃标记参数
跳跃目的核参地址参数
窗口计算提前结束标记参数

图2

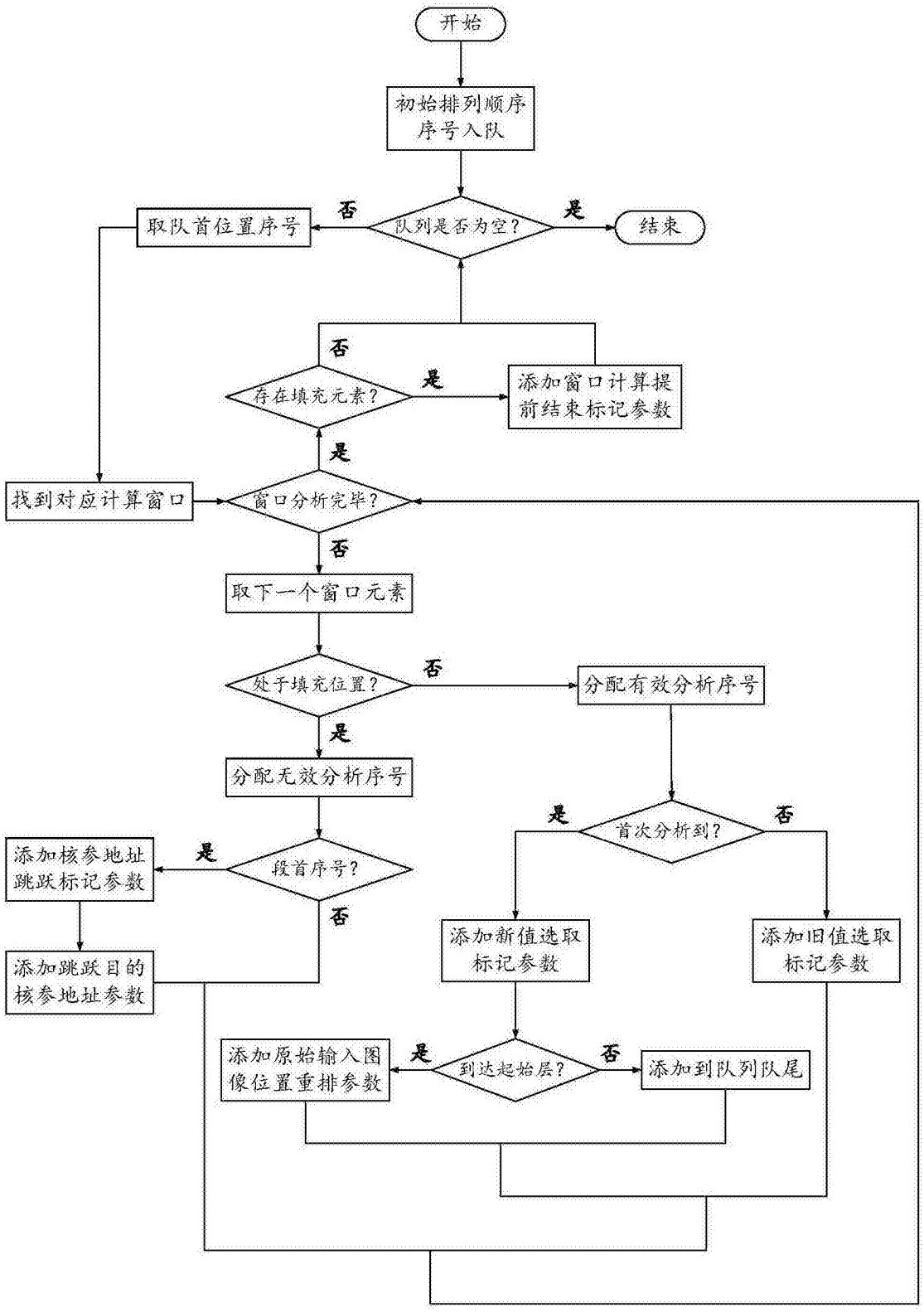

开始
初始排列顺序序号入队
队列是否为空?
否
是
取队首位置序号
结束
存在填充元素?
否
是
添加窗口计算提前结束标记参数
找到对应计算窗口
窗口分析完毕?
是
否
取下一个窗口元素
处于填充位置?
否
是
分配有效分析序号
分配无效分析序号
首次分析到?
是
否
段首序号?
是
否
添加核参地址跳跃标记参数
添加跳跃目的核参地址参数
添加新值选取标记参数
添加旧值选取标记参数
到达起始层?
是
否
添加原始输入图像位置重排参数
添加到队列队尾

图3

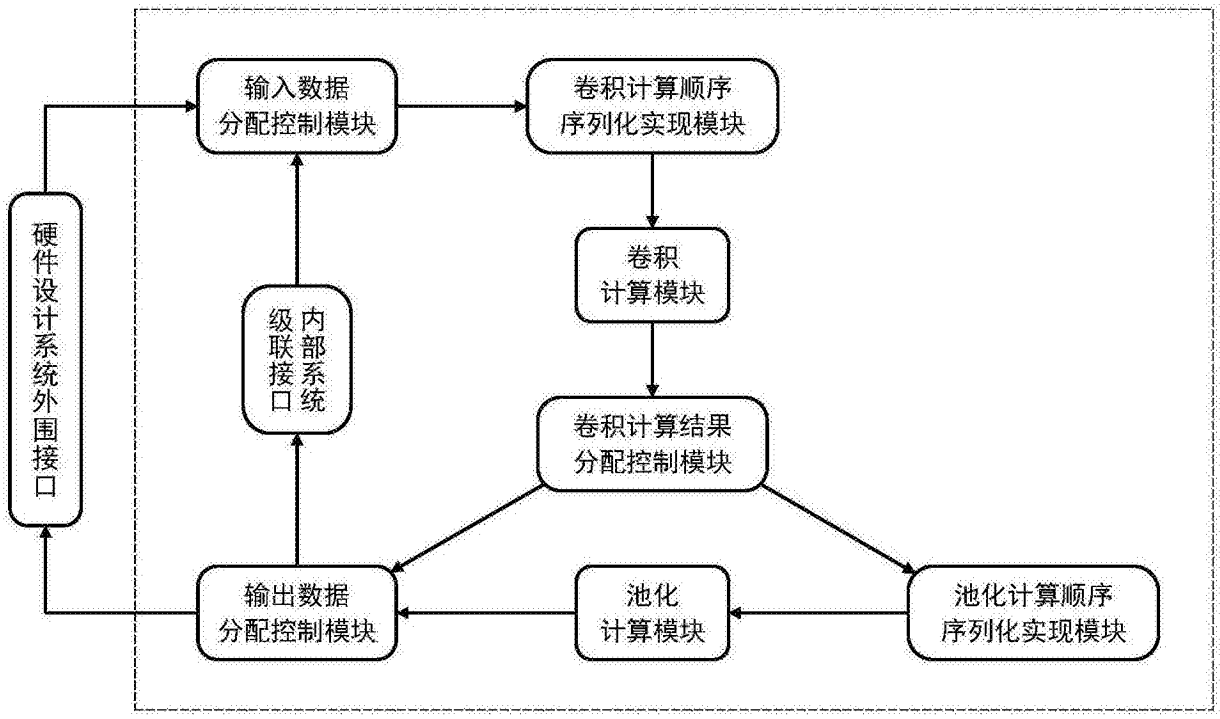

输入数据
分配控制模块
卷积计算顺序
序列化实现模块
卷积
计算模块
卷积计算结果
分配控制模块
池化计算顺序
序列化实现模块
池化
计算模块
输出数据
分配控制模块
级联接口
内部系统
硬件设计系统外围接口

图4

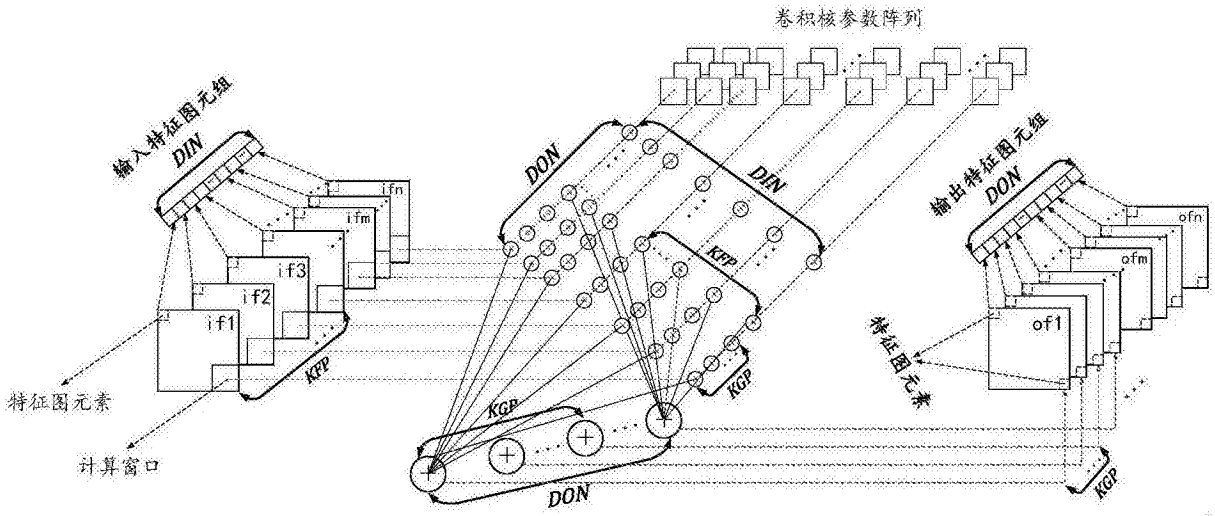

卷积核参数阵列
输入特征图元组
DIN
ifn
ifm
if3
if2
if1
KFP
特征图元素
计算窗口
DON
DIN
KFP
KGP
KGP
DON
输出特征图元组
DON
ofn
ofm
of1
特征图元素
KGP

图5

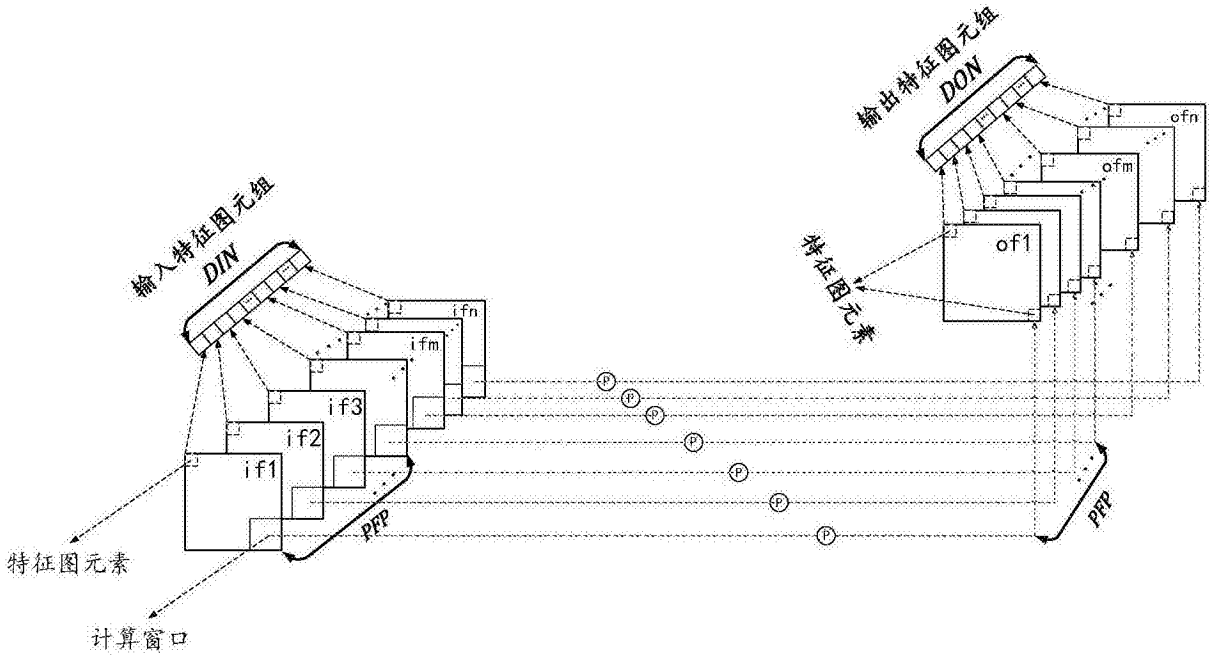

输入特征图元组
DIN
ifn
ifm
if3
if2
if1
PFP
特征图元素
计算窗口
输出特征图元组
DON
ofn
ofm
of1
特征图元素
PFP

图6

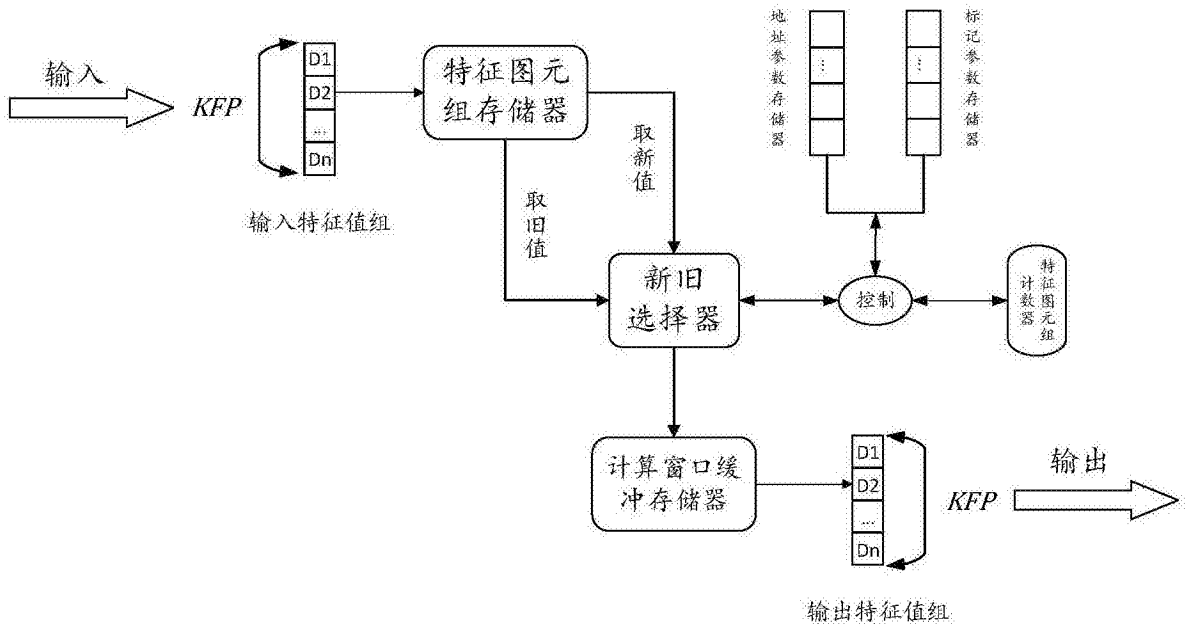

输入
KFP
D1
D2
...
Dn
输入特征值组
特征图元组存储器
取新值
取旧值
地址参数存储器
标记参数存储器
新旧选择器
控制
特征图元计数器
计算窗口缓冲存储器
D1
D2
...
Dn
KFP
输出
输出特征值组

图7

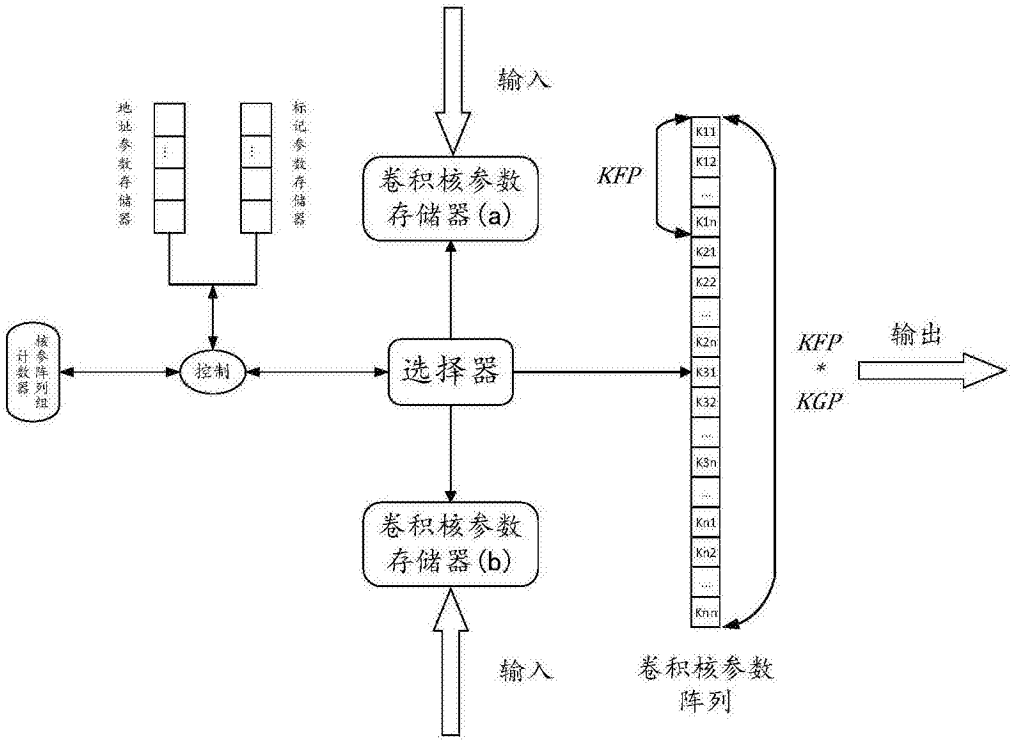

输入
地址参数存储器
标记参数存储器
卷积核参数存储器(a)
KFP
K11
K12
K1n
K21
K22
K2n
K31
K32
K3n
Kn1
Kn2
Knn
核参阵列组计数器
控制
选择器
KFP * KGP
输出
卷积核参数存储器(b)
输入
卷积核参数阵列

图8

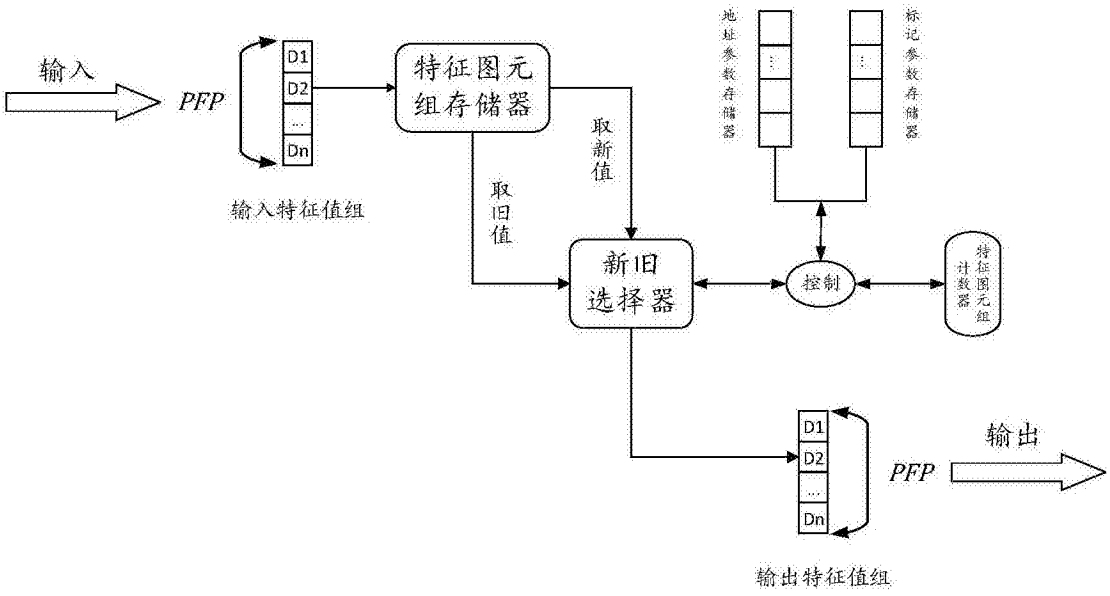

输入
PFP
D1
D2
Dn
输入特征值组
特征图元组存储器
取新值
取旧值
地址参数存储器
标记参数存储器
新旧选择器
控制
特征图元组计数器
D1
D2
Dn
PFP
输出
输出特征值组

图9

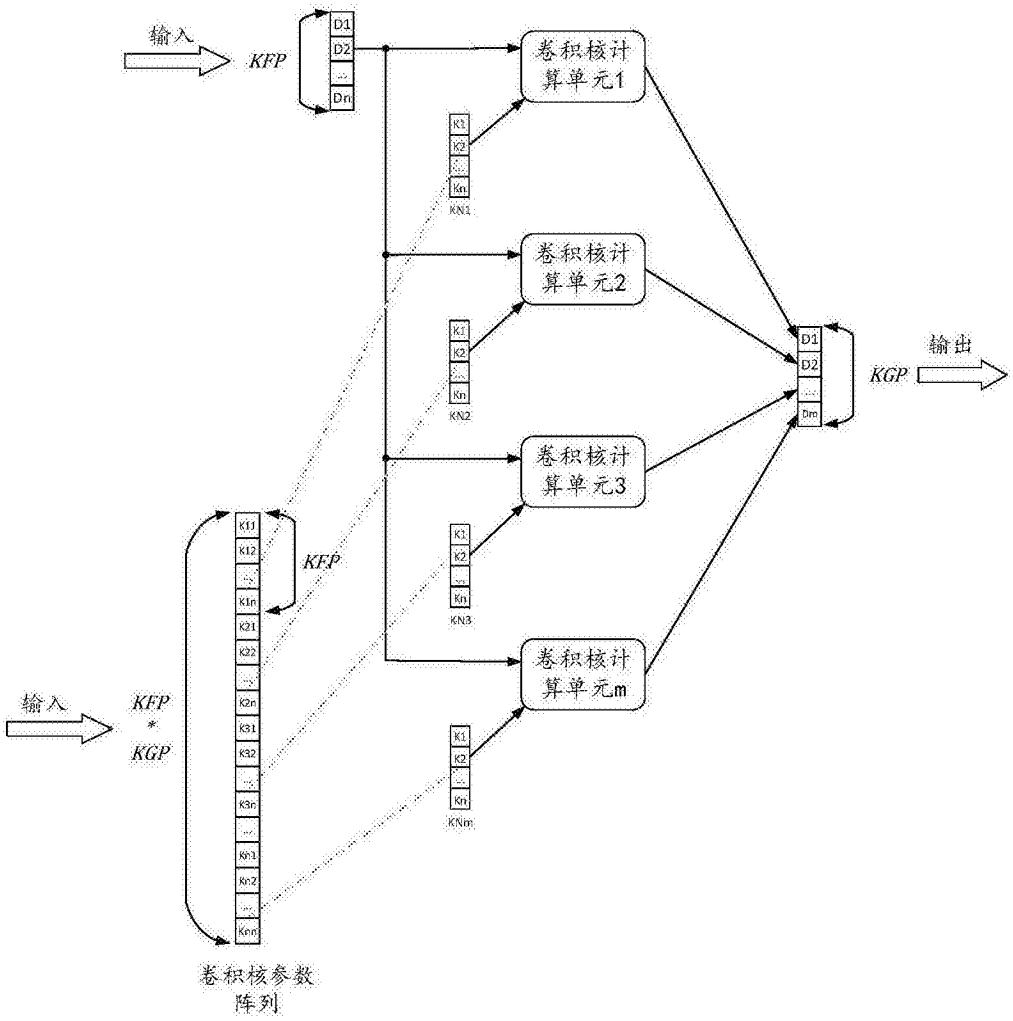

输入
KFP
D1
D2
Dn
卷积核计算单元1
卷积核计算单元2
卷积核计算单元3
卷积核计算单元m
K1
K2
Kn
KN1
KN2
KN3
KNm
D1
D2
Dm
输出
KGP
K11
K12
K1n
K21
K22
K2n
K31
K32
K3n
Kn1
Kn2
Knn
输入
KFP
*
KGP
卷积核参数阵列

图10

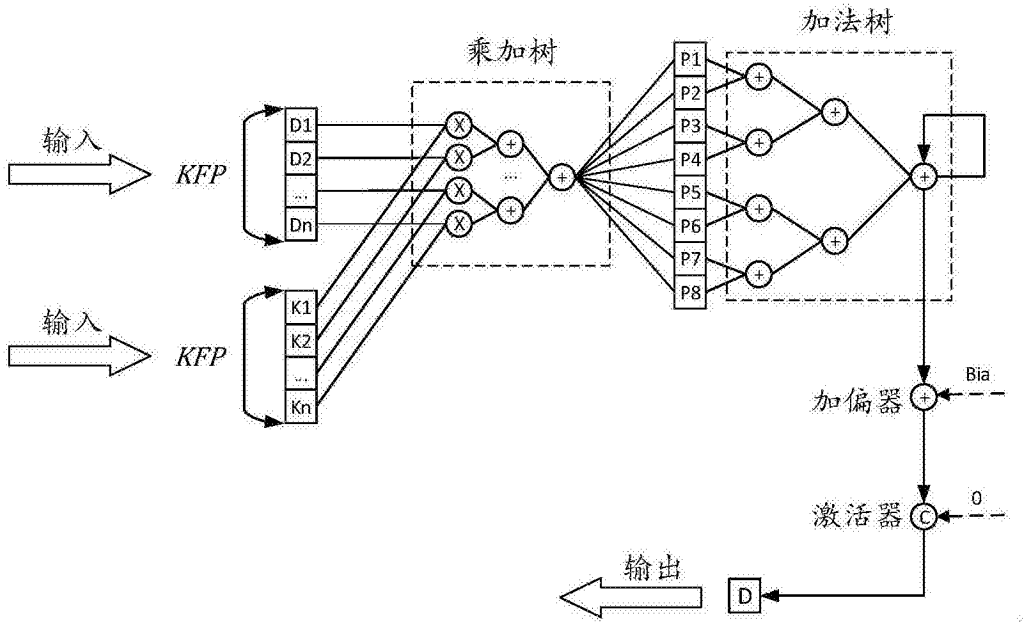

加法树
乘加树
输入
KFP
D1
D2
...
Dn
输入
KFP
K1
K2
...
Kn
P1
P2
P3
P4
P5
P6
P7
P8
加偏器
Bia
激活器
0
输出
D

图11

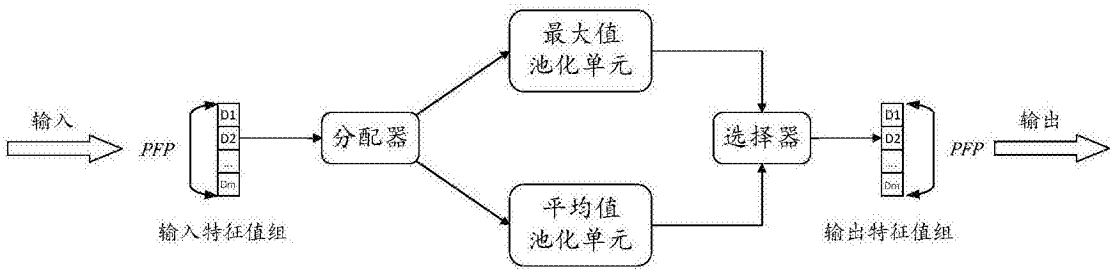

输入
PFP
D1
D2
...
Dn
输入特征值组
分配器
最大值
池化单元
平均值
池化单元
选择器
D1
D2
...
Dm
输出特征值组
PFP
输出

图12

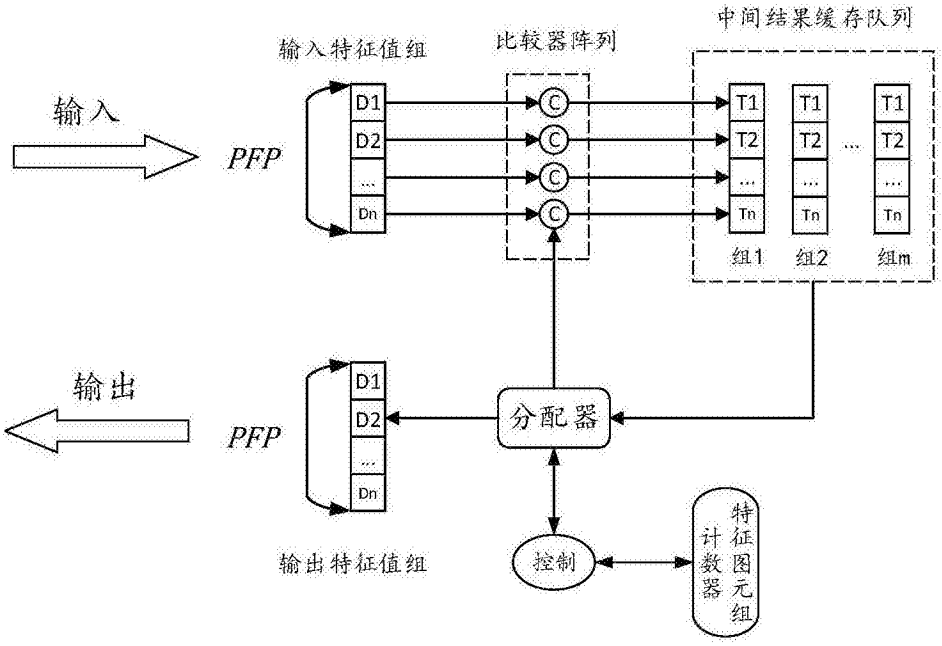

中间结果缓存队列
输入特征值组
比较器阵列
输入
PFP
D1
D2
...
Dn
C
T1
T2
Tn
组1
组2
组m
输出
分配器
控制
特征图元组计数器
输出特征值组

图13

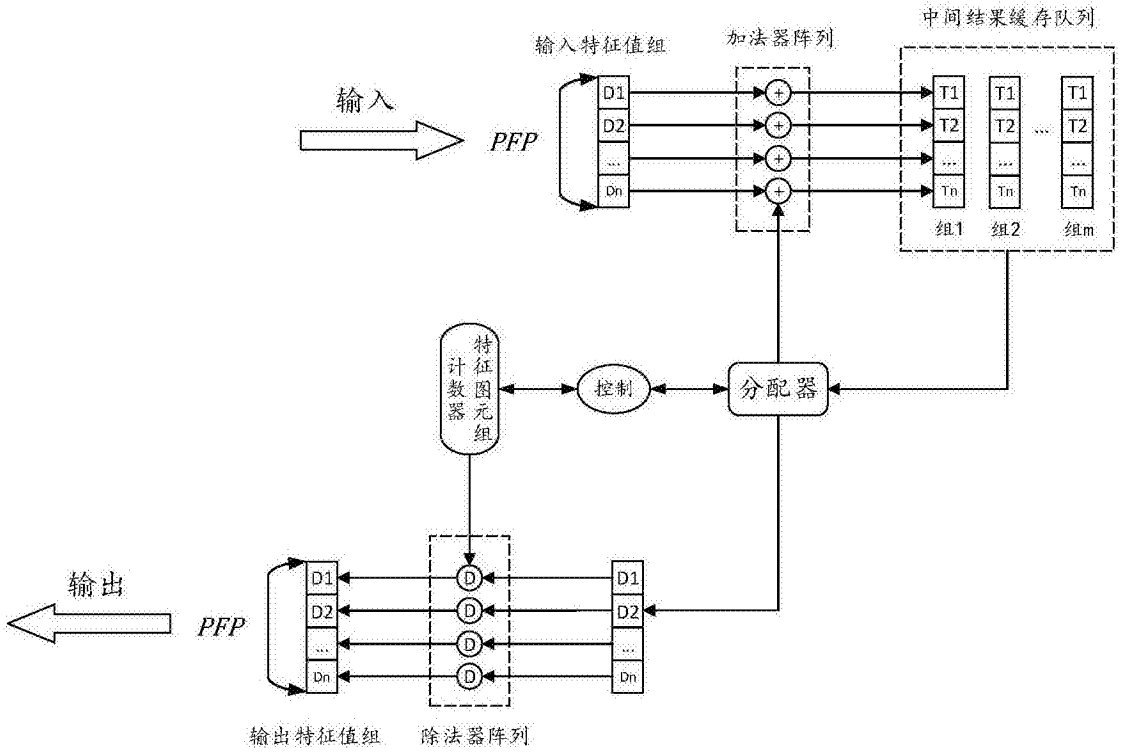

中间结果缓存队列
输入特征值组
加法器阵列
输入
PFP
D1
D2
...
Dn
T1
T2
Tn
组1
组2
组m
特征图元组计数器
控制
分配器
输出
D
输出特征值组
除法器阵列

图14