

# (12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织  
国际局

(43) 国际公布日  
2020 年 8 月 6 日 (06.08.2020)



(10) 国际公布号  
**WO 2020/155044 A1**

(51) 国际专利分类号:  
**G06N 3/063** (2006.01)

(21) 国际申请号: PCT/CN2019/074249

(22) 国际申请日: 2019 年 1 月 31 日 (31.01.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(71) 申请人: 深圳市大疆创新科技有限公司 (SZ DJI TECHNOLOGY CO., LTD.) [CN/CN]; 中国广东省深圳市南山区高新区南区粤兴一道9号香港科大深圳产学研大楼6楼, Guangdong 518057 (CN)。

(72) 发明人: 韩峰(HAN, Feng); 中国广东省深圳市南山区高新区南区粤兴一道9号香港科大深圳产学研大楼6楼, Guangdong 518057 (CN)。 杨康(YANG, Kang); 中国广东省深圳市南山区高新区南区粤兴一道9号香港科大深圳产学研大楼6楼, Guangdong 518057 (CN)。 谷骞(GU, Qian); 中国广东省深圳市南山区高新区南区粤兴一道9号香港科大深圳产学研大楼6楼, Guangdong 518057 (CN)。

(74) 代理人: 北京龙双利达知识产权代理有限公司 (LONGSUN LEAD IP LTD.); 中国北京市海淀区北清路68号院3号楼101, Beijing 100094 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国(除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT,

(54) Title: CONVOLUTION CALCULATION DEVICE AND METHOD, PROCESSOR AND MOVABLE DEVICE

(54) 发明名称: 卷积计算的装置、方法、处理器和可移动设备

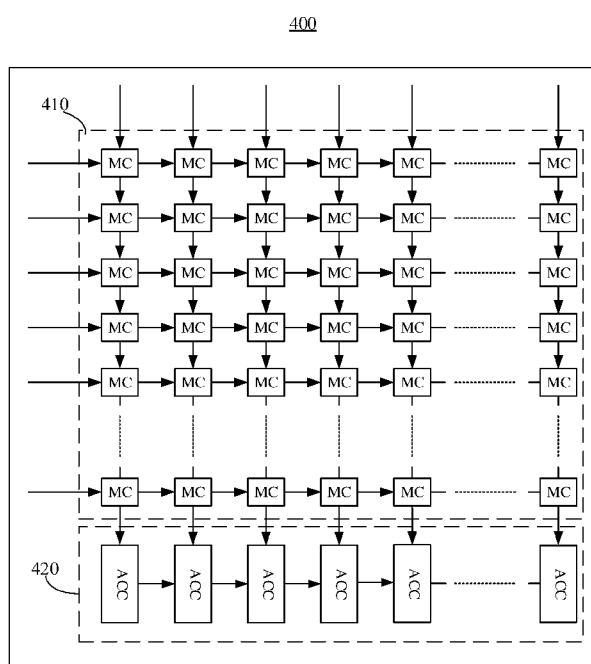


图4

(57) Abstract: A convolution calculation device (400) and method, a processor and a movable device. The device (400) comprises: an array of multiply-add units (410), comprising M rows and N columns of multiply-add units, a specific multiply-add unit among the M rows and N columns of multiply-add units being used for multiplying a feature value inputted into the specific multiply-add unit and a weight value corresponding to the specific multiply-add unit, adding the resulting product and a previous multiply-add output result, and outputting the resulting sum as an output result of the specific multiply-add unit; and an array of accumulate units (420), comprising one row of N accumulate units, the N accumulate units respectively corresponding to N columns of the array of multiply-add units (410), and a specific accumulate unit among the N accumulate units being used for adding an output result of the last multiply-add unit of a column corresponding to the specific accumulate unit to a previous accumulate output result, and outputting the resulting sum as an output result of the specific accumulate unit. The described technical solution may improve the efficiency of convolution calculation.

RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI,  
CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布：

— 包括国际检索报告(条约第21条(3))。

---

**(57) 摘要：**一种卷积计算的装置(400)、方法、处理器和可移动设备。该装置(400)包括：乘加单元阵列(410)，包括M行N列的乘加单元，所述M行N列的乘加单元中的特定乘加单元用于对输入到所述特定乘加单元的特征值和所述特定乘加单元对应的权重值进行相乘，并将相乘后的积与前一乘加输出结果相加，并将相加后的和作为所述特定乘加单元的输出结果输出；累加单元阵列(420)，包括1行N个累加单元，所述N个累加单元分别与所述乘加单元阵列(410)的N列对应，所述N个累加单元中的特定累加单元用于将所述特定累加单元对应列的最后一个乘加单元的输出结果与前一累加输出结果相加，并将相加后的和作为所述特定累加单元的输出结果输出。通过上述技术方案，能够提高卷积计算的效率。

## 卷积计算的装置、方法、处理器和可移动设备

## 版权申明

[0001] 本专利文件披露的内容包含受版权保护的材料。该版权为版权所有人所有。版权所有人不反对任何人复制专利与商标局的官方记录和档案中所存在的该专利文件或者该专利披露。

## 技术领域

[0002] 本申请涉及信息技术领域，并且更具体地，涉及一种卷积计算的装置、方法、处理器和可移动设备。

## 背景技术

[0003] 卷积神经网络（Convolutional Neural Network, CNN）是一种机器学习算法，它被广泛应用于目标识别、目标检测以及图像的语义分割等计算机视觉任务。

[0004] 目前卷积神经网络的卷积计算过程中，或者数据处理的并行度较低，或者需要较多的数据搬移操作，导致效率较低。因此，如何提高卷积计算的效率，成为卷积神经网络设计中一个亟待解决的技术问题。

## 20 发明内容

[0005] 本申请实施例提供了一种卷积计算的装置、方法、处理器和可移动设备，能够提高卷积计算的效率。

[0006] 第一方面，提供了一种卷积计算的装置，包括：乘加单元阵列，包括 M 行 N 列的乘加单元，所述 M 行 N 列的乘加单元中的特定乘加单元用于对输入到所述特定乘加单元的特征值和所述特定乘加单元对应的权重值进行相乘，并将相乘后的积与前一乘加输出结果相加，并将相加后的和作为所述特定乘加单元的输出结果输出，其中，所述特定乘加单元为所述 M 行 N 列的乘加单元中的任一乘加单元，所述前一乘加输出结果为所述特定乘加单元所在列中所述特定乘加单元的前一个乘加单元的输出结果或零，M 和 N 均为正整数；累加单元阵列，包括 1 行 N 个累加单元，所述 N 个累加单元分别与所述乘加单元阵列的 N 列对应，所述 N 个累加单元中的特定累加单元

用于将所述特定累加单元对应列的最后一个乘加单元的输出结果与前一累加输出结果相加，并将相加后的和作为所述特定累加单元的输出结果输出，其中，所述前一累加输出结果为所述特定累加单元的前一个累加单元的输出结果或零。

- 5 [0007] 第二方面，提供了一种卷积计算的方法，包括：向乘加单元阵列输入权重值，其中，所述乘加单元阵列包括  $M$  行  $N$  列的乘加单元；向所述乘加单元阵列输入特征值；通过所述  $M$  行  $N$  列的乘加单元中的特定乘加单元对输入到所述特定乘加单元的特征值和所述特定乘加单元对应的权重值进行相乘，并将相乘后的积与前一乘加输出结果相加，并将相加后的和作为所述
- 10 特定乘加单元的输出结果输出，其中，所述特定乘加单元为所述  $M$  行  $N$  列的乘加单元中的任一乘加单元，所述前一乘加输出结果为所述特定乘加单元所在列中所述特定乘加单元的前一个乘加单元的输出结果或零， $M$  和  $N$  均为正整数；通过累加单元阵列中的特定累加单元将所述特定累加单元对应列的最后一个乘加单元的输出结果与前一累加输出结果相加，并将相加后的和
- 15 作为所述特定累加单元的输出结果输出，其中，所述累加单元阵列包括 1 行  $N$  个累加单元，所述  $N$  个累加单元分别与所述乘加单元阵列的  $N$  列对应，所述前一累加输出结果为所述特定累加单元的前一个累加单元的输出结果或零。

[0008] 第三方面，提供了一种处理器，包括第一方面的卷积计算的装置。

- 20 [0009] 第四方面，提供了一种可移动设备，包括第一方面的卷积计算的装置；或者，第三方面的处理器。

[0010] 第五方面，提供了一种计算机存储介质，该计算机存储介质中存储有程序代码，该程序代码可以用于指示执行上述第二方面的方法。

- 25 [0011] 本申请实施例的技术方案，可以利用一个乘加单元完成该乘加单元对应的权重值的所有计算，减少了计算过程中的数据搬移，降低了输入和输出数据的带宽，从而能够提高卷积计算的效率。

附图说明

[0012] 图 1 是本申请实施例的卷积神经网络的卷积操作过程的示意图。

- 30 [0013] 图 2 是应用本申请实施例的技术方案的架构图。

[0014] 图 3 是本申请实施例的可移动设备的示意性架构图。

[0015] 图 4 是本申请一个实施例的卷积计算的装置的示意图。

[0016] 图 5 是本申请另一个实施例的卷积计算的装置的示意图。

[0017] 图 6 是本申请实施例的卷积核映射的示意图。

[0018] 图 7 是本申请实施例的乘加单元的示意图。

5 [0019] 图 8 是本申请实施例的累加单元的示意图。

[0020] 图 9 是本申请实施例的卷积计算的方法的示意性流程图。

### 具体实施方式

[0021] 下面将结合附图，对本申请实施例中的技术方案进行描述。

10 [0022] 应理解，本文中的具体的例子只是为了帮助本领域技术人员更好地理解本申请实施例，而非限制本申请实施例的范围。

[0023] 还应理解，在本申请的各种实施例中，各过程的序号的大小并不意味着执行顺序的先后，各过程的执行顺序应以其功能和内在逻辑确定，而不应对本申请实施例的实施过程构成任何限定。

15 [0024] 还应理解，本说明书中描述的各种实施方式，既可以单独实施，也可以组合实施，本申请实施例对此并不限定。

[0025] 本申请实施例的技术方案可以应用于各种深度学习算法，例如卷积神经网络，但本申请实施例对此并不限定。

[0026] 图 1 示出了卷积神经网络的卷积操作过程的示意图。

20 [0027] 如图 1 所示，卷积神经网络的卷积操作会对输入的一组权重值和一组输入特征图（Input Feature Map，IFM）进行运算后输出一组输出特征图（Output Feature Map，OFM）。输入的权重值被称为滤波器（Filter）或卷积核。输入特征图为上一层的输出特征图。输出特征图为输入特征图经过当前层运算后得到的特征图。卷积核和输入、输出特征图都可以被表示为一个多维矩阵，卷积神经网络的卷积层的一次卷积运算为输入特征矩阵的至少部分特征值（数据单元）与卷积核矩阵的权重值进行内积操作。

25 [0028] 卷积层的卷积运算可以采用滑动窗口方式，以输入特征值矩阵的左上角为起点，以卷积核的大小为窗口，依次滑动窗口到输入特征矩阵的右下角，产生一个完整的二维输出特征矩阵。每次滑动窗口后，卷积计算装置都会从输入特征值矩阵中提取一个窗口大小的输入特征值，将其与卷积核进行内积操作，产生一个输出特征值。依照上述方式，依次产生所有的二维输出特征

30

矩阵后，便可得到该卷积层的三维输出特征矩阵。

[0029] 图 2 是应用本申请实施例的技术方案的架构图。

[0030] 如图 2 所示，系统 200 可以包括卷积计算装置 210 和存储器 220。

5 [0031] 存储器 220 用于存储待处理的数据，例如，输入特征图和权重值，以及存储处理后的数据，例如输出特征图。存储器 220 可以为静态随机存取存储器（Static Random Access Memory，SRAM）。

[0032] 卷积计算装置 210 包括乘累加单元（Multiply Accumulate Unit，MAU）211、IFM 输入模块 212、权重值输入模块 213 和 OFM 存储模块 214。权重值输入模块 213 负责从存储器 220 中读出权重值，并按特定格式送到 MAU 211。IFM 输入模块 212 负责将输入特征图数据从存储器 220 中读出，并将其送到 MAU 211 中进行卷积运算。MAU 211 可以包括脉动阵列和用来存储中间计算结果的缓存。进行卷积运算时，MAU 211 首先将权重值输入模块 213 送入的权重值装载到脉动阵列，之后，当输入特征图数据从 IFM 输入模块 212 送入脉动阵列后，将其与提前装载的权重值进行乘累加。如果 MAU 211 15 中的缓存中缓存了中间结果，则脉动阵列输出结果还会继续与该缓存中的中间结果再进行一次乘累加。乘累加的结果如果仍为卷积运算的中间结果，则将其存储到 MAU 的缓存中，否则输出到下级模块 OFM 存储模块 214 中进行后续处理。OFM 存储模块 214 将 MAU 211 输出的卷积计算结果组装为存储器 220 中存储的数据格式，然后将其写入存储器 220。

20 [0033] 在本申请实施例中，利用卷积神经网络权重值共享的特性，减少计算过程中的数据搬移，进而降低输入和输出数据的带宽，从而提高卷积计算的效率。

[0034] 在一些实施例中，本申请实施例的技术方案可以应用于可移动设备中。该可移动设备可以是无人机、无人驾驶船、自动驾驶车辆或机器人等，本申请 25 实施例对此并不限定。

[0035] 图 3 是本申请实施例的可移动设备 300 的示意性架构图。

[0036] 如图 3 所示，可移动设备 300 可以包括动力系统 310、控制系统 320、传感系统 330 和处理系统 340。

[0037] 动力系统 310 用于为该可移动设备 300 提供动力。

30 [0038] 以无人机为例，无人机的动力系统可以包括电子调速器（简称为电调）、螺旋桨以及与螺旋桨相对应的电机。电机连接在电子调速器与螺旋桨之间，

电机和螺旋桨设置在对应的机臂上；电子调速器用于接收控制系统产生的驱动信号，并根据驱动信号提供驱动电流给电机，以控制电机的转速。电机用于驱动螺旋桨旋转，从而为无人机的飞行提供动力。

[0039] 传感系统 330 可以用于测量可移动设备 300 的姿态信息，即可移动设备 300 在空间的位置信息和状态信息，例如，三维位置、三维角度、三维速度、三维加速度和三维角速度等。传感系统 330 例如可以包括陀螺仪、电子罗盘、惯性测量单元（Inertial Measurement Unit, IMU）、视觉传感器、全球定位系统（Global Positioning System, GPS）、气压计、空速计等传感器中的至少一种。

10 [0040] 传感系统 330 还可用于采集图像，即传感系统 330 包括用于采集图像的传感器，例如相机等。

[0041] 控制系统 320 用于控制可移动设备 300 的移动。控制系统 320 可以按照预先设置的程序指令对可移动设备 300 进行控制。例如，控制系统 320 可以根据传感系统 330 测量的可移动设备 300 的姿态信息控制可移动设备 300 的移动。控制系统 320 也可以根据来自遥控器的控制信号对可移动设备 300 15 进行控制。例如，对于无人机，控制系统 320 可以为飞行控制系统（飞控），或者为飞控中的控制电路。

[0042] 处理系统 340 可以处理传感系统 330 采集的图像。例如，处理系统 340 可以为图像信号处理（Image Signal Processing, ISP）类芯片。

20 [0043] 处理系统 340 可以为图 2 中的系统 200，或者，处理系统 340 可以包括图 2 中的系统 200。

[0044] 应理解，上述对于可移动设备 300 的各组成部件的划分和命名仅仅是示例性的，并不应理解为对本申请实施例的限制。

25 [0045] 还应理解，可移动设备 300 还可以包括图 3 中未示出的其他部件，本申请实施例对此并不限定。

[0046] 图 4 示出了本申请一个实施例的卷积计算的装置 400 的示意图。该装置 400 可以为图 2 中的 MAU 211。

[0047] 如图 4 所示，该装置 400 可以包括乘加单元阵列 410 和累加单元阵列 420。

30 [0048] 乘加单元阵列 410，包括 M 行 N 列的乘加单元（Multiply AccumulateCell, MC），M 和 N 均为正整数。

[0049] 所述 M 行 N 列的乘加单元中的特定乘加单元用于对输入到所述特定乘加单元的特征值和所述特定乘加单元对应的权重值进行相乘，并将相乘后的积与前一乘加输出结果相加，并将相加后的和作为所述特定乘加单元的输

5 出结果输出，其中，所述特定乘加单元为所述 M 行 N 列的乘加单元中的任一乘加单元，所述前一乘加输出结果为所述特定乘加单元所在列中所述特定乘加单元的前一个乘加单元的输

[0050] 例如，对于第一行的乘加单元，前一乘加输出结果为零，该乘加单元将相应特征值和权重值相乘后向下输出，即输出到所在列的下一个乘加单元。对于其他行的乘加单元，前一乘加输出结果为所在列的前一个乘加单元的输

10 出结果，该乘加单元将相应特征值和权重值相乘后，再与前一乘加输出结果相加，将相加后的和向下输出。其中，对于最后一行的乘加单元，输出给所在列的累加单元，对于其他乘加单元，输出给所在列的下一个乘加单元。

[0051] 累加单元阵列，包括 1 行 N 个累加单元 (AccumulateCell, ACC)，所述 N 个累加单元分别与所述乘加单元阵列的 N 列对应，所述 N 个累加单元中的特定累加单元用于将所述特定累加单元对应列的最后一个乘加单元的输

15 出结果与前一累加输出结果相加，并将相加后的和作为所述特定累加单元的输

[0052] 例如，若某一累加单元为卷积核对应的第一个累加单元，则前一累加输出结果为零，该累加单元只需将所在列的最后一个乘加单元的输

20 出结果输出给下一个累加单元。若某一累加单元不是卷积核对应的第一个累加单元，则前一累加输出结果为该累加单元的前一个累加单元的输

[0053] 应理解，本申请实施例中的各种单元或模块具体可以由电路实现，例如，乘加单元可以为乘加电路，但本申请实施例对此并不限定，它们也可以由其他方式实现。

25

[0054] 可选地，在本申请一个实施例中，如图 5 所示，该装置 400 还可以包括：权重值注入模块 430 和特征值注入模块 440。

[0055] 权重值注入模块 430，用于向所述乘加单元阵列 410 输入权重值；

30

[0056] 特征值注入模块 440，用于向所述乘加单元阵列 410 输入特征值。

[0057] 所述权重值注入模块 440 与所述乘加单元阵列 410 的第一行乘加单元连接, 其中, 对于每一列乘加单元, 权重值从第一行的乘加单元传递到对应的乘加单元。每一列乘加单元在权重值传递到对应的乘加单元后, 同时将权重值锁存。

- 5 [0058] 具体而言, 权重值注入模块 430 与乘加单元阵列 410 的每一列乘加单元仅有一个接口 (即权重值注入模块 430 与第一行乘加单元之间的接口), 该接口每个时钟周期仅能传输一个权重值。权重输入可分为移位和装载两个阶段。在移位阶段, 权重值注入模块 430 依次将同一列乘加单元需要的权重值通过同一个接口依次送入乘加单元阵列 410。在乘加单元阵列 410 中, 接收到的权重值从接口处的乘加单元依次向下传递。在装载阶段, 脉动阵列中同一列的乘加单元同时将权重值装载到各自的缓存中。权重值注入模块 430 为相邻两列乘加单元输入权重值时会有一个时钟周期的延迟。

- 15 [0059] 所述特征值注入模块 440 与所述乘加单元阵列 410 的第一列乘加单元连接, 其中对于每一行乘加单元, 特征值从第一列的乘加单元依次向下一列的乘加单元传递。

- [0060] 具体而言, 特征值注入模块 440 与乘加单元阵列 410 的每一行乘加单元仅有一个接口 (即特征值注入模块 440 与第一列乘加单元之间的接口), 该接口每个时钟周期仅能传输一个特征值。在乘加单元阵列 410 中, 接收到的特征值从接口处的乘加单元依次向右传递直至最后一个乘加单元。特征值注入模块 440 为相邻两行乘加单元输入特征值时会有一个时钟周期的延迟。

[0061] 可选地, 在本申请一个实施例中, 所述乘加单元阵列 410 可以按照如下方式与卷积核的权重值对应:

- 25 [0062] 对于  $K_d \times K_h \times K_w$  大小的卷积核, 在深度方向展开的  $K_d$  个  $K_h \times K_w$  的二维图, 沿所述乘加单元阵列的列方向映射到对应的乘加单元上, 其中, 一个权重值对应一个乘加单元,  $K_d$ 、 $K_h$  和  $K_w$  均为正整数, 分别表示卷积核的深度、高度和宽度;

[0063] 多个卷积核沿所述乘加单元阵列的行方向进行映射。

[0064] 可选地, 在  $K_d \times K_h > M$  的情况下, 将一个卷积核分多次映射到所述乘加单元阵列上。

- 30 [0065] 例如, 图 6 示出了卷积核映射的示意图。图 6 中  $K_h$  和  $K_w$  分别为 3。卷积核 0 和卷积核 1 分别展开为  $K_d$  个  $3 \times 3$  的二维图, 然后卷积核 0 展开得

到的二维图沿着乘加单元阵列的 0~2 列乘加单元依次映射，卷积核 1 展开得到的二维图沿着脉动阵列的 3~5 列乘加单元依次映射。假如乘加单元阵列的高度 M 等于  $Kd \times 3$ ，则卷积核 0 的第一个二维图映射在脉动阵列第 0~2 列的第 0~2 行的乘加单元，第 Kd 个二维图映射到第 0~2 列的最后三行乘加单元。卷积核 1 的第一个二维图映射在脉动阵列第 3~5 列的第 0~2 行的乘加单元，第 Kd 个二维图映射到第 3~5 列的最后三行乘加单元。

[0066] 累加单元阵列的映射方式与上方的乘加单元阵列的映射方式类似。如图 6 所示，卷积核 0 映射到第 0~2 列的累加单元，卷积核 1 映射到第 3~5 列的累加单元。其中卷积核 0 的计算结果从第 2 列的累加单元输出，卷积核 1 的计算结果从第 5 列的累加单元输出。

[0067] 图 7 示出了本申请一个实施例的乘加单元的示意图。

[0068] 如图 7 所示，特定乘加单元可以包括：

[0069] 权重值移位寄存器 701，用于缓存以及沿所述特定乘加单元所在列传递权重值，并将所述特定乘加单元对应的权重值锁存到权重值寄存器 702；

[0070] 权重值寄存器 702，用于缓存所述特定乘加单元对应的权重值；

[0071] 特征值移位寄存器 703，用于缓存以及沿所述特定乘加单元所在行传递特征值，并将特征值锁存到特征值寄存器 704；

[0072] 特征值寄存器 704，用于缓存特征值；

[0073] 相乘电路 705，用于对所述权重值寄存器 702 中权重值和所述特征值寄存器 704 中的特征值进行相乘，并将相乘后的积输出到积寄存器 706；

[0074] 积寄存器 706，用于缓存所述相乘电路 705 相乘后的积；

[0075] 相加电路 707，用于将所述积寄存器 706 中的积与所述前一乘加输出结果相加，并将相加后的和沿所述特定乘加单元所在列向下输出。

[0076] 具体而言，权重值移位寄存器 701 负责缓存从权重值注入模块 430 或上一个乘加单元送来的权重值。在权重值输入的移位阶段，权重值移位寄存器 701 缓存的权重值会向下传递到下一个乘加单元。在权重输入的装载阶段，权重值移位寄存器 701 缓存的权重值会被锁存到权重值寄存器 702。特征值移位寄存器 703 负责缓存从特征值注入模块 440 或左边的乘加单元送来的特征值数据。特征值移位寄存器 703 缓存的特征值数据会被锁存到特征值寄存器 704，同时还会被送到右边的乘加单元。相乘电路 705 负责将权重值寄存器 702 和特征值寄存器 704 缓存的权重值和特征值进行乘法操作，操作结果

被送到积寄存器 706 中。相加电路 707 负责将积寄存器 706 中的数据和上面送入的前一乘加输出结果再次累加后向下传递。

[0077] 图 8 示出了本申请一个实施例的累加单元的示意图。

[0078] 如图 8 所示, 特定累加单元可以包括:

- 5 [0079] 过滤电路 801, 用于根据卷积计算的步长值, 过滤所述特定累加单元对应列的最后一个乘加单元的输出结果, 并将过滤后的结果输出到乘加单元结果寄存器 802;

[0080] 乘加单元结果寄存器 802, 用于缓存所述过滤电路 801 过滤后的结果;

- 10 [0081] 延迟电路 803, 用于根据卷积计算的膨胀值对所述前一累加输出结果进行延迟, 并将延迟后的结果输出到累加单元结果寄存器 804;

[0082] 累加单元结果寄存器 804, 用于缓存所述延迟电路 803 延迟后的结果;

[0083] 第一阶段相加电路 805, 用于将所述乘加单元结果寄存器 802 中的结果与所述累加单元结果寄存器 804 中的结果相加, 并将相加后的和输出到和寄存器 806;

- 15 [0084] 和寄存器 806, 用于缓存所述第一阶段相加电路 805 相加后的和。

[0085] 可选地, 所述特定累加单元还可以包括: 第二阶段相加电路 807。

- [0086] 在这种情况下, 所述和寄存器 806, 用于在所述特定累加单元对应的列不为特定卷积核对应的最后一列时, 将所述和寄存器 806 中的和输出到下一个累加单元; 在所述特定累加单元对应的列为所述特定卷积核对应的最后一列时, 将所述和寄存器 806 中的和输出到所述第二阶段相加电路 807;
- 20

[0087] 所述第二阶段相加电路 807, 用于在所述特定累加单元对应的列为所述特定卷积核对应的最后一列时, 将所述和寄存器 806 中的和与中间结果缓存模块中的所述特定卷积核的中间结果相加, 并将相加后的和输出到结果处理模块。

- 25 [0088] 具体而言, 乘加单元计算的结果中会存在冗余的结果(无效的结果)。以图 6 中第 1 列乘加单元为例, 该列乘加单元中的权重值与特征值矩阵中的第 1 列的特征值的乘积为有效的结果, 而与特征值矩阵中的第 0 列的特征值的乘积为冗余的结果, 需要过滤掉。因此, 过滤电路 801 可以根据卷积计算时输入的参数步长(Stride)值过滤掉乘加单元阵列 410 输出的冗余的结果,
- 30 同时, 将过滤后的结果送入乘加单元结果寄存器 802。

[0089] 另一方面, 每列乘加单元输出的有效的结果需要与相应的另一列输出

的有效的结果累加。而这两列输出有效的结果的时间可能间隔一些时钟周期。因此，延迟电路 803 需要将左边累加单元输出的结果延迟指定时钟周期后送入累加单元结果寄存器 804。延迟的时钟周期数由卷积计算时输入的参数膨胀 (Dilation) 值计算得到。

5 [0090] 第一阶段相加电路 805 负责将乘加单元结果寄存器 802 和累加单元结果寄存器 804 中缓存的数据进行累加后送入和寄存器 806。

[0091] 在将卷积运算的卷积核映射时，连续  $K_w$  个累加单元会映射到同一个卷积核， $K_w$  的大小和卷积核的宽度相同。 $K_w$  个累加单元中，第一个累加单元不需要接收左边累加单元输出的结果，同时，最后一个累加单元也不会  
10 将和寄存器 806 缓存的结果输出到右边的乘加单元，它只会在第二阶段相加电路 807 中将和寄存器 806 缓存的结果与从中间结果缓存模块中读回的中间结果累加后输出到结果处理模块。

[0092] 可选地，在本申请一个实施例中，如图 5 所示，该装置 400 还可以包括：结果处理模块 450。

15 [0093] 结果处理模块 450，用于处理所述累加单元阵列 410 输出的结果。

[0094] 可选地，该装置 400 还可以包括：中间结果缓存模块 460。

[0095] 在这种情况下，所述结果处理模块 450，在所述累加单元阵列 410 输出的结果为特定卷积核的最终结果时，将所述累加单元阵列 410 输出的结果输出；在所述累加单元阵列 410 输出的结果为所述特定卷积核的中间结果时，  
20 将所述累加单元阵列 410 输出的结果缓存到所述中间结果缓存模块 460。

[0096] 具体而言，在  $K_d \times K_h > M$  的情况下，一个卷积核分多次映射到所述乘加单元阵列 410 上。在这种情况下，该卷积核的计算结果由多次映射得到的计算结果组成。也就是说，多次映射中的前几次映射得到的计算结果为中间结果，要与下一次映射得到的计算结果相加，最终得到该卷积核的最终结果。因此，结果处理模块 450 在累加单元阵列 410 输出的结果为特定卷积核  
25 的中间结果时，将输出的结果缓存到中间结果缓存模块 460，以便于与下一次输出的结果相加；在累加单元阵列 410 输出的结果为特定卷积核的最终结果时，将累加单元阵列 410 输出的结果输出。

[0097] 中间结果缓存模块 460 用于缓存每个卷积核的中间结果。可选地，中间结果缓存模块可以包括：N 个先入先出队列 (First Input First Output, FIFO)。  
30

[0098] 对于  $K_d \times K_h \times K_w$  大小的卷积核，所述 N 个 FIFO 中的每  $K_w$  个 FIFO

组成一组，用于缓存一个卷积核的中间结果。

[0099]  $K_w$  个 FIFO 组成一组缓存一个卷积核的中间结果，可以充分利用  $K_w$  个 FIFO 的资源，从而可以提高 FIFO 的利用率。

5 [00100] 可选地，在本申请一个实施例中，如图 5 所示，该装置 400 还可以包括：控制模块 470。

[00101] 控制模块 470 可用于控制权重值和特征值到所述乘加单元阵列 410 的输入，以及控制所述乘加单元阵列 410 和所述累加单元阵列 420 的计算。

10 [00102] 具体而言，控制模块 470 可用于控制装置 400 中各模块的处理，以得到计算结果。例如，控制模块 470 可以首先控制权重值注入模块 430 将权重值输入模块送入的权重值装载到乘加单元阵列 410，然后控制特征值注入模块 440 将 IFM 输入模块送入的特征值输入乘加单元阵列 410，并控制乘加单元阵列 410 和累加单元阵列 420 进行卷积运算。待所有的特征图数据送入乘加单元阵列 410 并完成卷积运算后，再依次重复上述过程，直至完成所有  
15 的卷积运算。

[00103] 本申请实施例的技术方案，可以利用一个乘加单元完成该乘加单元对应的权重值的所有计算，减少了计算过程中的数据搬移，降低了输入和输出数据的带宽，从而能够提高卷积计算的效率。

20 [00104] 以上描述了本申请实施例的卷积计算的装置，下面描述本申请实施例的卷积计算的方法。本申请实施例的卷积计算的方法为前述本申请实施例的卷积计算的装置或包括本申请实施例的卷积计算的装置的设备实施本申请实施例的技术方案时的方法，相关描述可以参考前述实施例，以下为了简洁，在此不再赘述。

[00105] 图 9 示出了本申请实施例的卷积计算的方法 900 的示意性流程图。

25 [00106] 如图 9 所示，所述方法 900 包括：

[00107] 910，向乘加单元阵列输入权重值，其中，所述乘加单元阵列包括  $M$  行  $N$  列的乘加单元；

[00108] 920，向所述乘加单元阵列输入特征值；

30 [00109] 930，通过所述  $M$  行  $N$  列的乘加单元中的特定乘加单元对输入到所述特定乘加单元的特征值和所述特定乘加单元对应的权重值进行相乘，并将相乘后的积与前一乘加输出结果相加，并将相加后的和作为所述特定乘加

单元的输出结果输出, 其中, 所述特定乘加单元为所述 M 行 N 列的乘加单元中的任一乘加单元, 所述前一乘加输出结果为所述特定乘加单元所在列中所述特定乘加单元的前一个乘加单元的输出结果或零, M 和 N 均为正整数;

[00110] 940, 通过累加单元阵列中的特定累加单元将所述特定累加单元对应列的最后一个乘加单元的输出结果与前一累加输出结果相加, 并将相加后的和作为所述特定累加单元的输出结果输出, 其中, 所述累加单元阵列包括 1 行 N 个累加单元, 所述 N 个累加单元分别与所述乘加单元阵列的 N 列对应, 所述前一累加输出结果为所述特定累加单元的前一个累加单元的输出结果或零。

10 [00111] 可选地, 在本申请一个实施例中, 所述向乘加单元阵列输入权重值, 包括: 通过所述乘加单元阵列的第一行乘加单元输入权重值, 其中, 对于每一列乘加单元, 权重值从第一行的乘加单元传递到对应的乘加单元。

[00112] 可选地, 在本申请一个实施例中, 在权重值传递到对应的乘加单元后, 每一列乘加单元同时将权重值锁存。

15 [00113] 可选地, 在本申请一个实施例中, 所述向所述乘加单元阵列输入特征值, 包括: 通过所述乘加单元阵列的第一列乘加单元输入特征值, 其中对于每一行乘加单元, 特征值从第一列的乘加单元依次向下一列的乘加单元传递。

20 [00114] 可选地, 在本申请一个实施例中, 所述乘加单元阵列按照如下方式与卷积核的权重值对应:

[00115] 对于  $K_d \times K_h \times K_w$  大小的卷积核, 在深度方向展开的  $K_d$  个  $K_h \times K_w$  的二维图, 沿所述乘加单元阵列的列方向映射到对应的乘加单元上, 其中, 一个权重值对应一个乘加单元,  $K_d$ 、 $K_h$  和  $K_w$  均为正整数, 分别表示卷积核的深度、高度和宽度;

25 [00116] 多个卷积核沿所述乘加单元阵列的行方向进行映射。

[00117] 可选地, 在本申请一个实施例中, 在  $K_d \times K_h > M$  的情况下, 将一个卷积核分多次映射到所述乘加单元阵列上。

[00118] 可选地, 在本申请一个实施例中, 所述方法还包括: 处理所述累加单元阵列输出的结果。

30 [00119] 可选地, 在本申请一个实施例中, 所述处理所述累加单元阵列输出的结果, 包括: 在所述累加单元阵列输出的结果为特定卷积核的最终结果

时，将所述累加单元阵列输出的结果输出；在所述累加单元阵列输出的结果为所述特定卷积核的中间结果时，将所述累加单元阵列输出的结果缓存到中间结果缓存模块。

[00120] 可选地，在本申请一个实施例中，所述中间结果缓存模块包括：N个先入先出队列 FIFO，其中，对于  $K_d \times K_h \times K_w$  大小的卷积核，所述 N 个 FIFO 中的每  $K_w$  个 FIFO 组成一组，用于缓存一个卷积核的中间结果。

[00121] 可选地，在本申请一个实施例中，所述特定乘加单元包括：权重值移位寄存器、权重值寄存器、特征值移位寄存器、特征值寄存器、相乘电路、积寄存器和相加电路；其中，通过所述权重值移位寄存器缓存以及沿所述特定乘加单元所在列传递权重值，并将所述特定乘加单元对应的权重值锁存到所述权重值寄存器；通过所述特征值移位寄存器缓存以及沿所述特定乘加单元所在行传递特征值，并将特征值锁存到所述特征值寄存器；通过所述相乘电路对所述权重值寄存器中权重值和所述特征值寄存器中的特征值进行相乘，并将相乘后的积输出到所述积寄存器；通过所述相加电路将所述积寄存器中的积与所述前一乘加输出结果相加，并将相加后的和沿所述特定乘加单元所在列向下输出。

[00122] 可选地，在本申请一个实施例中，所述特定累加单元包括：过滤电路、乘加单元结果寄存器、延迟电路、累加单元结果寄存器、第一阶段相加电路和和寄存器；其中，通过所述过滤电路，根据卷积计算的步长值，过滤所述特定累加单元对应列的最后一个乘加单元的输出结果，并将过滤后的结果输出到所述乘加单元结果寄存器；通过所述延迟电路，根据卷积计算的膨胀值对所述前一累加输出结果进行延迟，并将延迟后的结果输出到所述累加单元结果寄存器；通过所述第一阶段相加电路将所述乘加单元结果寄存器中的结果与所述累加单元结果寄存器中的结果相加，并将相加后的和输出到所述和寄存器。

[00123] 可选地，在本申请一个实施例中，所述特定累加单元还包括：第二阶段相加电路；其中，在所述特定累加单元对应的列不为特定卷积核对应的最后一列时，将所述和寄存器中的和输出到下一个累加单元；在所述特定累加单元对应的列为所述特定卷积核对应的最后一列时，将所述和寄存器中的和输出到所述第二阶段相加电路；通过所述第二阶段相加电路，在所述特定累加单元对应的列为所述特定卷积核对应的最后一列时，将所述和寄存器

中的和与中间结果缓存模块中的所述特定卷积核的中间结果相加。

[00124] 可选地, 在本申请一个实施例中, 所述方法还包括: 控制权重值和特征值到所述乘加单元阵列的输入, 以及控制所述乘加单元阵列和所述累加单元阵列的计算。

5 [00125] 本申请实施例还提供了一种处理器, 该处理器包括前述本申请实施例的卷积计算的装置。

[00126] 例如, 该处理器可以为图 2 中的卷积计算装置 210, 其中, MAU 211 可以为本申请实施例的卷积计算的装置。

10 [00127] 本申请实施例还提供了一种可移动设备, 该可移动设备可以包括上述本申请实施例的卷积计算的装置; 或者, 包括上述本申请实施例的处理器。

[00128] 本申请实施例还提供了一种计算机存储介质, 该计算机存储介质中存储有程序代码, 该程序代码可以用于指示执行上述本申请实施例的卷积计算的方法。

15 [00129] 本领域普通技术人员可以意识到, 结合本文中所公开的实施例描述的各示例的单元及算法步骤, 能够以电子硬件、计算机软件或者二者的结合来实现, 为了清楚地说明硬件和软件的可互换性, 在上述说明中已经按照功能一般性地描述了各示例的组成及步骤。这些功能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申请的范围。

20 [00130] 在本申请所提供的几个实施例中, 应该理解到, 所揭露的系统、装置和方法, 可以通过其它的方式实现。例如, 以上所描述的装置实施例仅仅是示意性的, 例如, 所述单元的划分, 仅仅为一种逻辑功能划分, 实际实现时可以有另外的划分方式, 例如多个单元或组件可以结合或者可以集成到另一个系统, 或一些特征可以忽略, 或不执行。另外, 所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口、装置或单元的间接耦合或通信连接, 也可以是电的, 机械的或其它的形式连接。

25 [00131] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的, 作为单元显示的部件可以是或者也可以不是物理单元, 即可以位于一个地方, 或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的

部分或者全部单元来实现本申请实施例方案的目的。

[00132] 另外，在本申请各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以是两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

[00133] 所述集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用，可以存储在一个计算机可读取存储介质中。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分，或者该技术方案的全部或部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备（可以是个人计算机，服务器，或者网络设备等）执行本申请各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：U 盘、移动硬盘、只读存储器（ROM，Read-Only Memory）、随机存取存储器（RAM，Random Access Memory）、磁碟或者光盘等各种可以存储程序代码的介质。

[00134] 以上所述，仅为本申请的具体实施方式，但本申请的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本申请揭露的技术范围内，可轻易想到各种等效的修改或替换，这些修改或替换都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应以权利要求的保护范围为准。

## 权利要求

1. 一种卷积计算的装置，其特征在于，包括：

乘加单元阵列，包括 M 行 N 列的乘加单元，所述 M 行 N 列的乘加单元中的特定乘加单元用于对输入到所述特定乘加单元的特征值和所述特定乘加单元对应的权重值进行相乘，并将相乘后的积与前一乘加输出结果相加，并将相加后的和作为所述特定乘加单元的输出结果输出，其中，所述特定乘加单元为所述 M 行 N 列的乘加单元中的任一乘加单元，所述前一乘加输出结果为所述特定乘加单元所在列中所述特定乘加单元的前一个乘加单元的输出结果或零，M 和 N 均为正整数；

累加单元阵列，包括 1 行 N 个累加单元，所述 N 个累加单元分别与所述乘加单元阵列的 N 列对应，所述 N 个累加单元中的特定累加单元用于将所述特定累加单元对应列的最后一个乘加单元的输出结果与前一累加输出结果相加，并将相加后的和作为所述特定累加单元的输出结果输出，其中，所述前一累加输出结果为所述特定累加单元的前一个累加单元的输出结果或零。

2. 根据权利要求 1 所述的装置，其特征在于，所述装置还包括：

权重值注入模块，用于向所述乘加单元阵列输入权重值；

特征值注入模块，用于向所述乘加单元阵列输入特征值。

3. 根据权利要求 2 所述的装置，其特征在于，所述权重值注入模块与所述乘加单元阵列的第一行乘加单元连接，其中，对于每一列乘加单元，权重值从第一行的乘加单元传递到对应的乘加单元。

4. 根据权利要求 3 所述的装置，其特征在于，每一列乘加单元用于在权重值传递到对应的乘加单元后，同时将权重值锁存。

5. 根据权利要求 2 至 4 中任一项所述的装置，其特征在于，所述特征值注入模块与所述乘加单元阵列的第一列乘加单元连接，其中对于每一行乘加单元，特征值从第一列的乘加单元依次向下一列的乘加单元传递。

6. 根据权利要求 1 至 5 中任一项所述的装置，其特征在于，所述乘加单元阵列用于按照如下方式与卷积核的权重值对应：

对于  $K_d \times K_h \times K_w$  大小的卷积核，在深度方向展开的  $K_d$  个  $K_h \times K_w$  的二维图，沿所述乘加单元阵列的列方向映射到对应的乘加单元上，其中，一个权重值对应一个乘加单元， $K_d$ 、 $K_h$  和  $K_w$  均为正整数，分别表示卷积核

的深度、高度和宽度;

多个卷积核沿所述乘加单元阵列的行方向进行映射。

7. 根据权利要求 6 所述的装置, 其特征在于, 在  $K_d \times K_h > M$  的情况下, 将一个卷积核分多次映射到所述乘加单元阵列上。

5        8. 根据权利要求 1 至 7 中任一项所述的装置, 其特征在于, 所述装置还包括:

结果处理模块, 用于处理所述累加单元阵列输出的结果。

9. 根据权利要求 8 所述的装置, 其特征在于, 所述装置还包括:

中间结果缓存模块;

10        其中, 所述结果处理模块用于, 在所述累加单元阵列输出的结果为特定卷积核的最终结果时, 将所述累加单元阵列输出的结果输出; 在所述累加单元阵列输出的结果为所述特定卷积核的中间结果时, 将所述累加单元阵列输出的结果缓存到所述中间结果缓存模块。

15        10. 根据权利要求 9 所述的装置, 其特征在于, 所述中间结果缓存模块包括:

N 个先入先出队列 FIFO, 其中, 对于  $K_d \times K_h \times K_w$  大小的卷积核, 所述 N 个 FIFO 中的每  $K_w$  个 FIFO 组成一组, 用于缓存一个卷积核的中间结果。

20        11. 根据权利要求 1 至 10 中任一项所述的装置, 其特征在于, 所述特定乘加单元包括:

权重值移位寄存器, 用于缓存以及沿所述特定乘加单元所在列传递权重值, 并将所述特定乘加单元对应的权重值锁存到权重值寄存器;

权重值寄存器, 用于缓存所述特定乘加单元对应的权重值;

25        特征值移位寄存器, 用于缓存以及沿所述特定乘加单元所在行传递特征值, 并将特征值锁存到特征值寄存器;

特征值寄存器, 用于缓存特征值;

相乘电路, 用于对所述权重值寄存器中权重值和所述特征值寄存器中的特征值进行相乘, 并将相乘后的积输出到积寄存器;

积寄存器, 用于缓存所述相乘电路相乘后的积;

30        相加电路, 用于将所述积寄存器中的积与所述前一乘加输出结果相加, 并将相加后的和沿所述特定乘加单元所在列向下输出。

12. 根据权利要求 1 至 11 中任一项所述的装置，其特征在于，所述特定累加单元包括：

5 过滤电路，用于根据卷积计算的步长值，过滤所述特定累加单元对应列的最后一个乘加单元的输出结果，并将过滤后的结果输出到乘加单元结果寄存器；

乘加单元结果寄存器，用于缓存所述过滤电路过滤后的结果；

延迟电路，用于根据卷积计算的膨胀值对所述前一累加输出结果进行延迟，并将延迟后的结果输出到累加单元结果寄存器；

累加单元结果寄存器，用于缓存所述延迟电路延迟后的结果；

10 第一阶段相加电路，用于将所述乘加单元结果寄存器中的结果与所述累加单元结果寄存器中的结果相加，并将相加后的和输出到和寄存器；

和寄存器，用于缓存所述第一阶段相加电路相加后的和。

13. 根据权利要求 12 所述的装置，其特征在于，所述特定累加单元还包括：

15 第二阶段相加电路；

其中，所述和寄存器，用于在所述特定累加单元对应的列不为特定卷积核对应的最后一列时，将所述和寄存器中的和输出到下一个累加单元；在所述特定累加单元对应的列为所述特定卷积核对应的最后一列时，将所述和寄存器中的和输出到所述第二阶段相加电路；

20 所述第二阶段相加电路，用于在所述特定累加单元对应的列为所述特定卷积核对应的最后一列时，将所述和寄存器中的和与中间结果缓存模块中的所述特定卷积核的中间结果相加，并将相加后的和输出到结果处理模块。

14. 根据权利要求 1 至 13 中任一项所述的装置，其特征在于，所述装置还包括：

25 控制模块，用于控制权重值和特征值到所述乘加单元阵列的输入，以及控制所述乘加单元阵列和所述累加单元阵列的计算。

15. 一种卷积计算的方法，其特征在于，包括：

向乘加单元阵列输入权重值，其中，所述乘加单元阵列包括  $M$  行  $N$  列的乘加单元；

30 向所述乘加单元阵列输入特征值；

通过所述  $M$  行  $N$  列的乘加单元中的特定乘加单元对输入到所述特定乘

加单元的特征值和所述特定乘加单元对应的权重值进行相乘，并将相乘后的积与前一乘加输出结果相加，并将相加后的和作为所述特定乘加单元的输  
结果输出，其中，所述特定乘加单元为所述  $M$  行  $N$  列的乘加单元中的任一  
乘加单元，所述前一乘加输出结果为所述特定乘加单元所在列中所述特定乘  
5 加单元的前一个乘加单元的输结果或零， $M$  和  $N$  均为正整数；

通过累加单元阵列中的特定累加单元将所述特定累加单元对应列的最  
后一个乘加单元的输结果与前一累加输结果相加，并将相加后的和作为  
所述特定累加单元的输结果输出，其中，所述累加单元阵列包括 1 行  $N$  个  
累加单元，所述  $N$  个累加单元分别与所述乘加单元阵列的  $N$  列对应，所述  
10 前一累加输结果为所述特定累加单元的前一个累加单元的输结果或零。

16. 根据权利要求 15 所述的方法，其特征在于，所述向乘加单元阵列  
输入权重值，包括：

通过所述乘加单元阵列的第一行乘加单元输入权重值，其中，对于每一  
列乘加单元，权重值从第一行的乘加单元传递到对应的乘加单元。

15 17. 根据权利要求 16 所述的方法，其特征在于，在权重值传递到对应  
的乘加单元后，每一列乘加单元同时将权重值锁存。

18. 根据权利要求 15 至 17 中任一项所述的方法，其特征在于，所述向  
所述乘加单元阵列输入特征值，包括：

20 通过所述乘加单元阵列的第一列乘加单元输入特征值，其中对于每一行  
乘加单元，特征值从第一列的乘加单元依次向下一列的乘加单元传递。

19. 根据权利要求 15 至 18 中任一项所述的方法，其特征在于，所述乘  
加单元阵列按照如下方式与卷积核的权重值对应：

25 对于  $K_d \times K_h \times K_w$  大小的卷积核，在深度方向展开的  $K_d$  个  $K_h \times K_w$  的  
二维图，沿所述乘加单元阵列的列方向映射到对应的乘加单元上，其中，一  
个权重值对应一个乘加单元， $K_d$ 、 $K_h$  和  $K_w$  均为正整数，分别表示卷积核  
的深度、高度和宽度；

多个卷积核沿所述乘加单元阵列的行方向进行映射。

20. 根据权利要求 19 所述的方法，其特征在于，在  $K_d \times K_h > M$  的情况  
下，将一个卷积核分多次映射到所述乘加单元阵列上。

30 21. 根据权利要求 15 至 20 中任一项所述的方法，其特征在于，所述方  
法还包括：

处理所述累加单元阵列输出的结果。

22. 根据权利要求 21 所述的方法，其特征在于，所述处理所述累加单元阵列输出的结果，包括：

5 在所述累加单元阵列输出的结果为特定卷积核的最终结果时，将所述累加单元阵列输出的结果输出；在所述累加单元阵列输出的结果为所述特定卷积核的中间结果时，将所述累加单元阵列输出的结果缓存到中间结果缓存模块。

23. 根据权利要求 22 所述的方法，其特征在于，所述中间结果缓存模块包括：

10 N 个先入先出队列 FIFO，其中，对于  $K_d \times K_h \times K_w$  大小的卷积核，所述 N 个 FIFO 中的每  $K_w$  个 FIFO 组成一组，用于缓存一个卷积核的中间结果。

24. 根据权利要求 15 至 23 中任一项所述的方法，其特征在于，所述特定乘加单元包括：权重值移位寄存器、权重值寄存器、特征值移位寄存器、特征值寄存器、相乘电路、积寄存器和相加电路；

其中，通过所述权重值移位寄存器缓存以及沿所述特定乘加单元所在列传递权重值，并将所述特定乘加单元对应的权重值锁存到所述权重值寄存器；

通过所述特征值移位寄存器缓存以及沿所述特定乘加单元所在行传递特征值，并将特征值锁存到所述特征值寄存器；

20 通过所述相乘电路对所述权重值寄存器中权重值和所述特征值寄存器中的特征值进行相乘，并将相乘后的积输出到所述积寄存器；

通过所述相加电路将所述积寄存器中的积与所述前一乘加输出结果相加，并将相加后的和沿所述特定乘加单元所在列向下输出。

25 25. 根据权利要求 15 至 24 中任一项所述的方法，其特征在于，所述特定累加单元包括：过滤电路、乘加单元结果寄存器、延迟电路、累加单元结果寄存器、第一阶段相加电路和和寄存器；

其中，通过所述过滤电路，根据卷积计算的步长值，过滤所述特定累加单元对应列的最后一个乘加单元的输出结果，并将过滤后的结果输出到所述乘加单元结果寄存器；

30 通过所述延迟电路，根据卷积计算的膨胀值对所述前一累加输出结果进行延迟，并将延迟后的结果输出到所述累加单元结果寄存器；

通过所述第一阶段相加电路将所述乘加单元结果寄存器中的结果与所述累加单元结果寄存器中的结果相加，并将相加后的和输出到所述和寄存器。

26. 根据权利要求 25 所述的方法，其特征在于，所述特定累加单元还包括：第二阶段相加电路；

5 其中，在所述特定累加单元对应的列不为特定卷积核对应的最后一列时，将所述和寄存器中的和输出到下一个累加单元；在所述特定累加单元对应的列为所述特定卷积核对应的最后一列时，将所述和寄存器中的和输出到所述第二阶段相加电路；

10 通过所述第二阶段相加电路，在所述特定累加单元对应的列为所述特定卷积核对应的最后一列时，将所述和寄存器中的和与中间结果缓存模块中的所述特定卷积核的中间结果相加。

27. 根据权利要求 15 至 26 中任一项所述的方法，其特征在于，所述方法还包括：

15 控制权重值和特征值到所述乘加单元阵列的输入，以及控制所述乘加单元阵列和所述累加单元阵列的计算。

28. 一种处理器，其特征在于，包括根据权利要求 1 至 14 中任一项所述的卷积计算的装置。

29. 一种可移动设备，其特征在于，包括：

20 根据权利要求 1 至 14 中任一项所述的卷积计算的装置；或者，  
根据权利要求 28 所述的处理器。

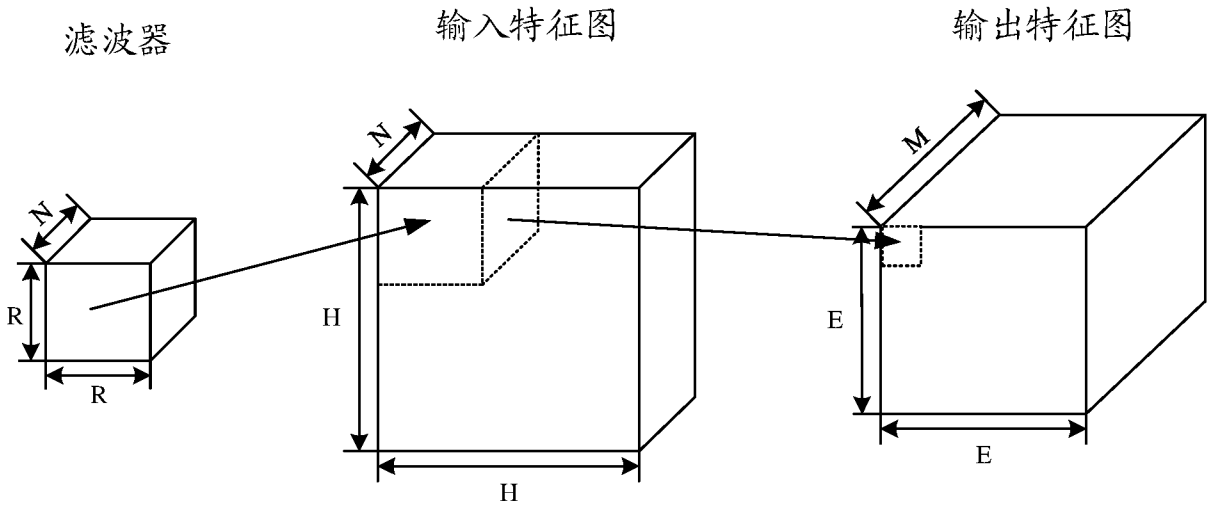


图1

200

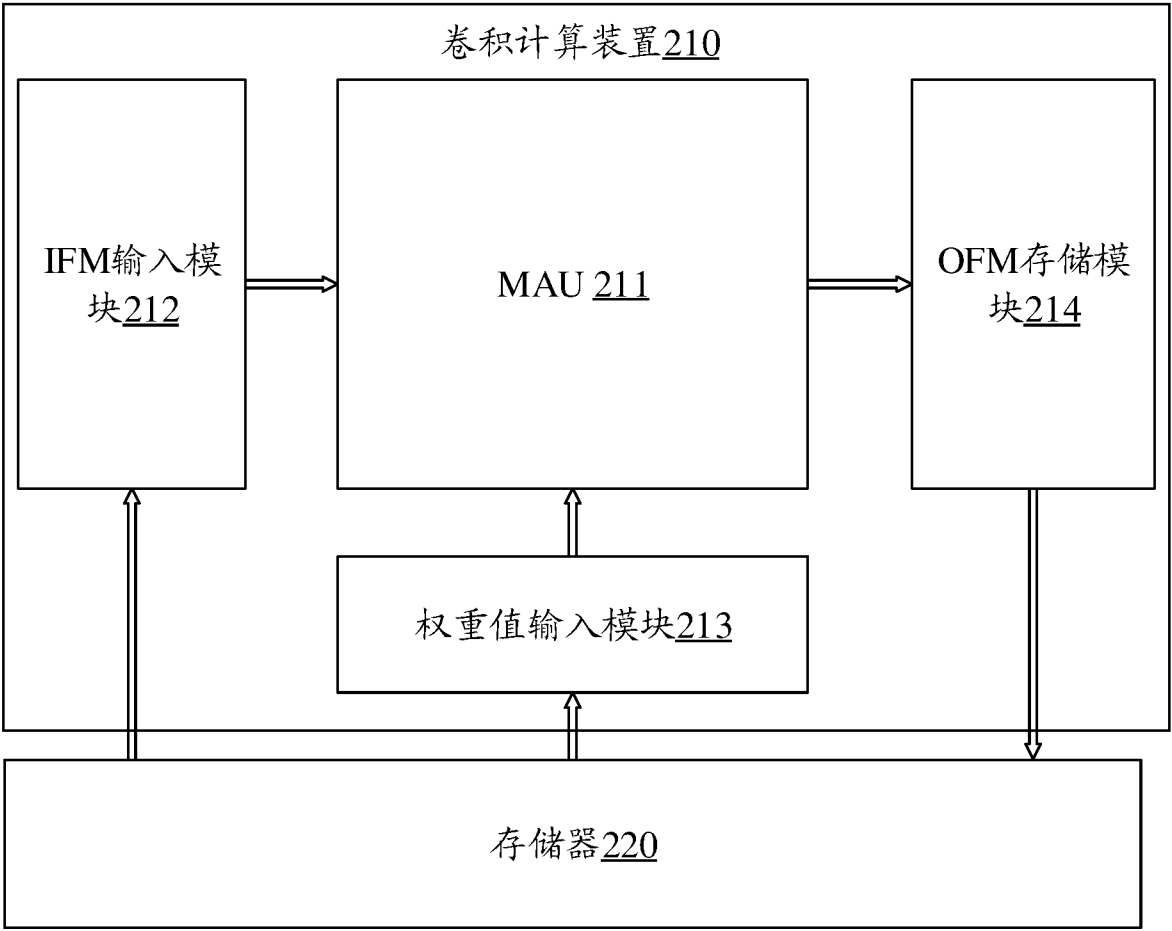


图2

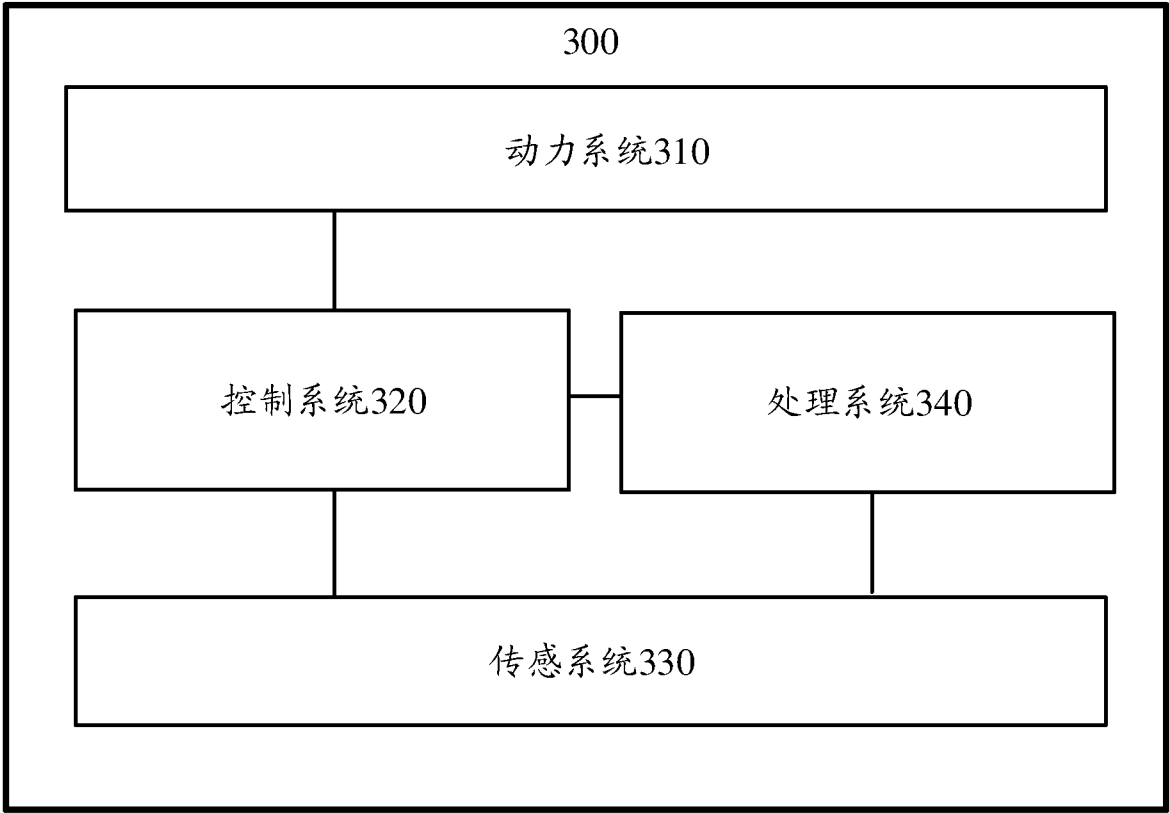


图3

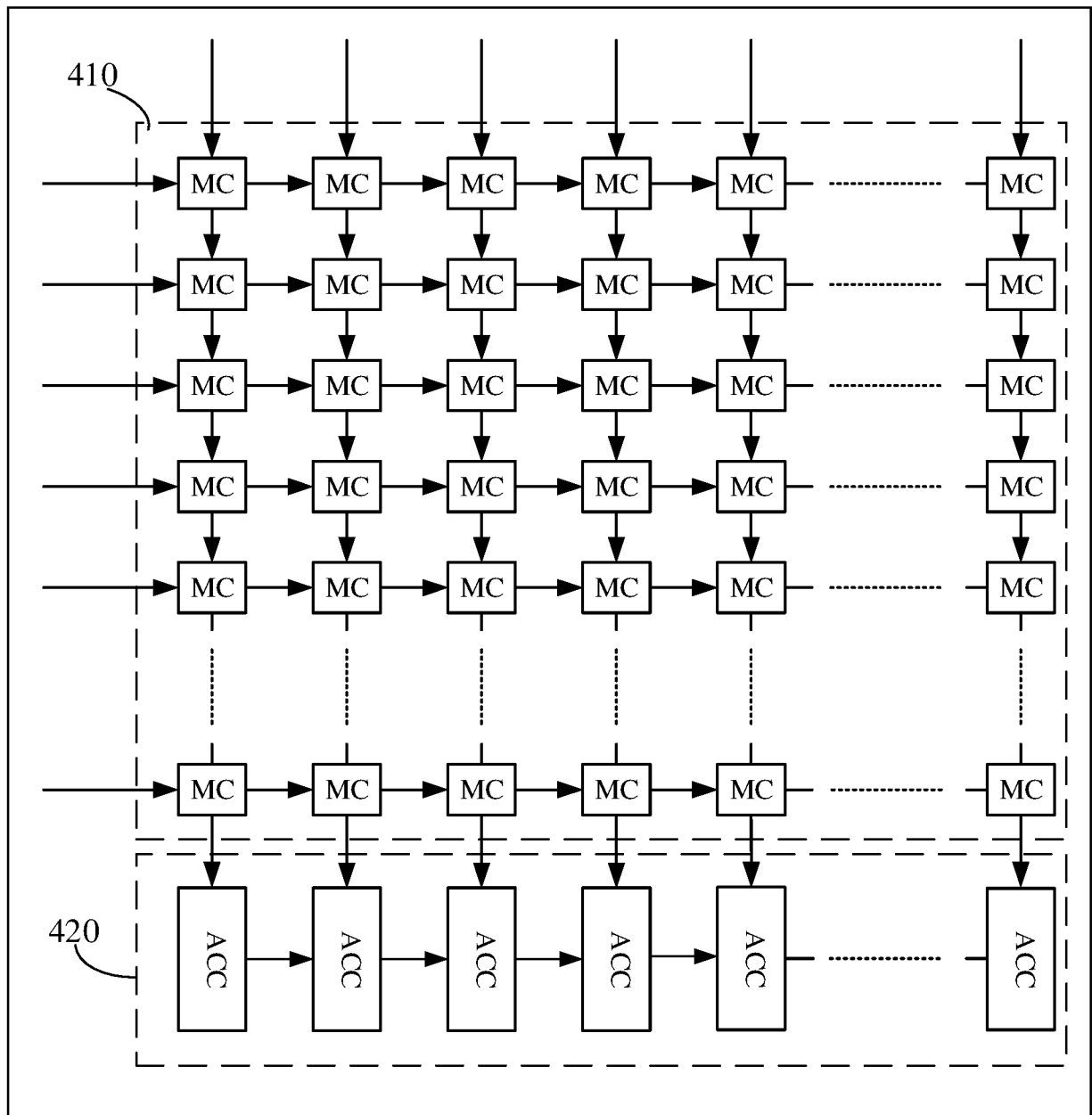
400

图4

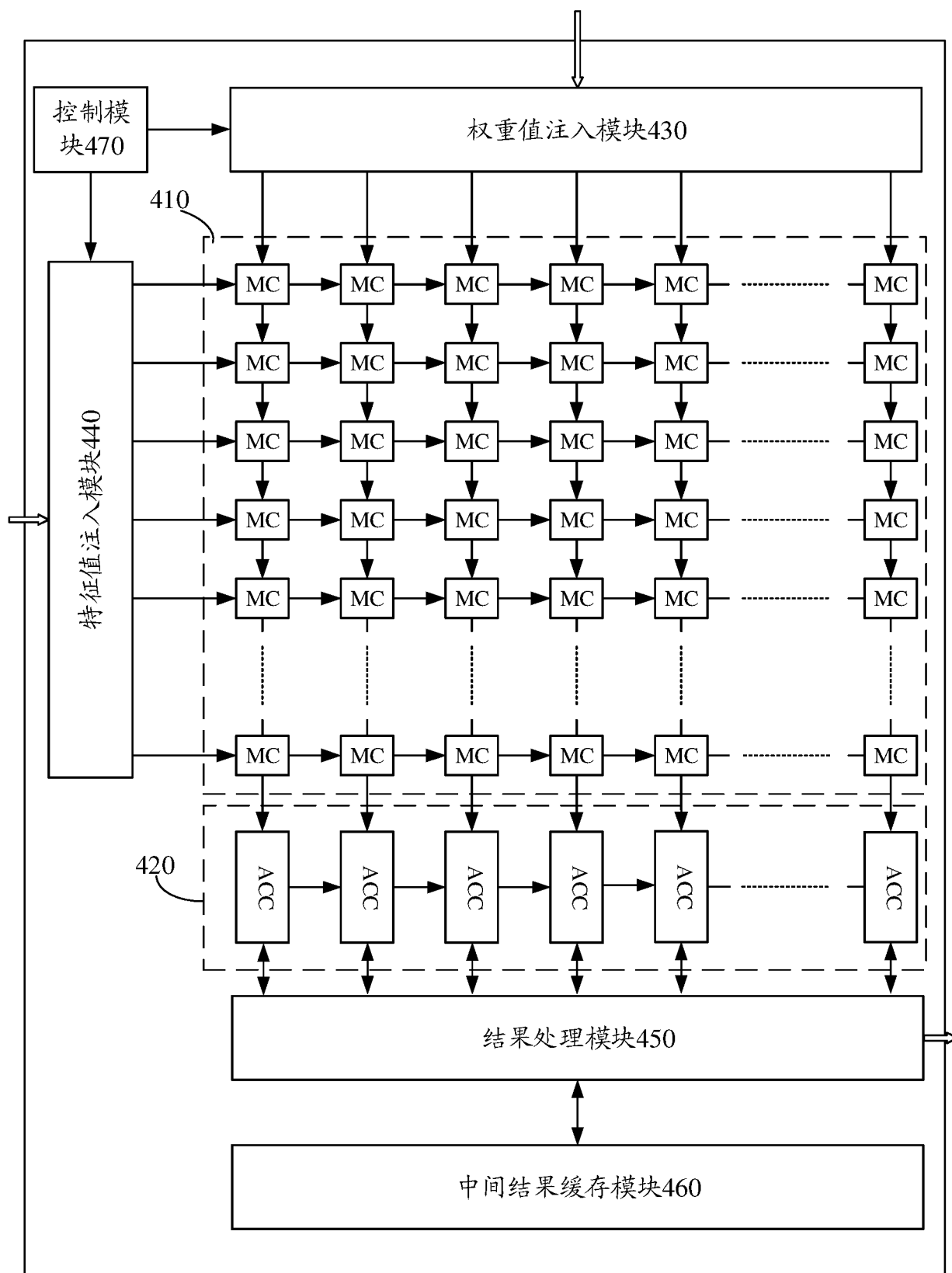


图5

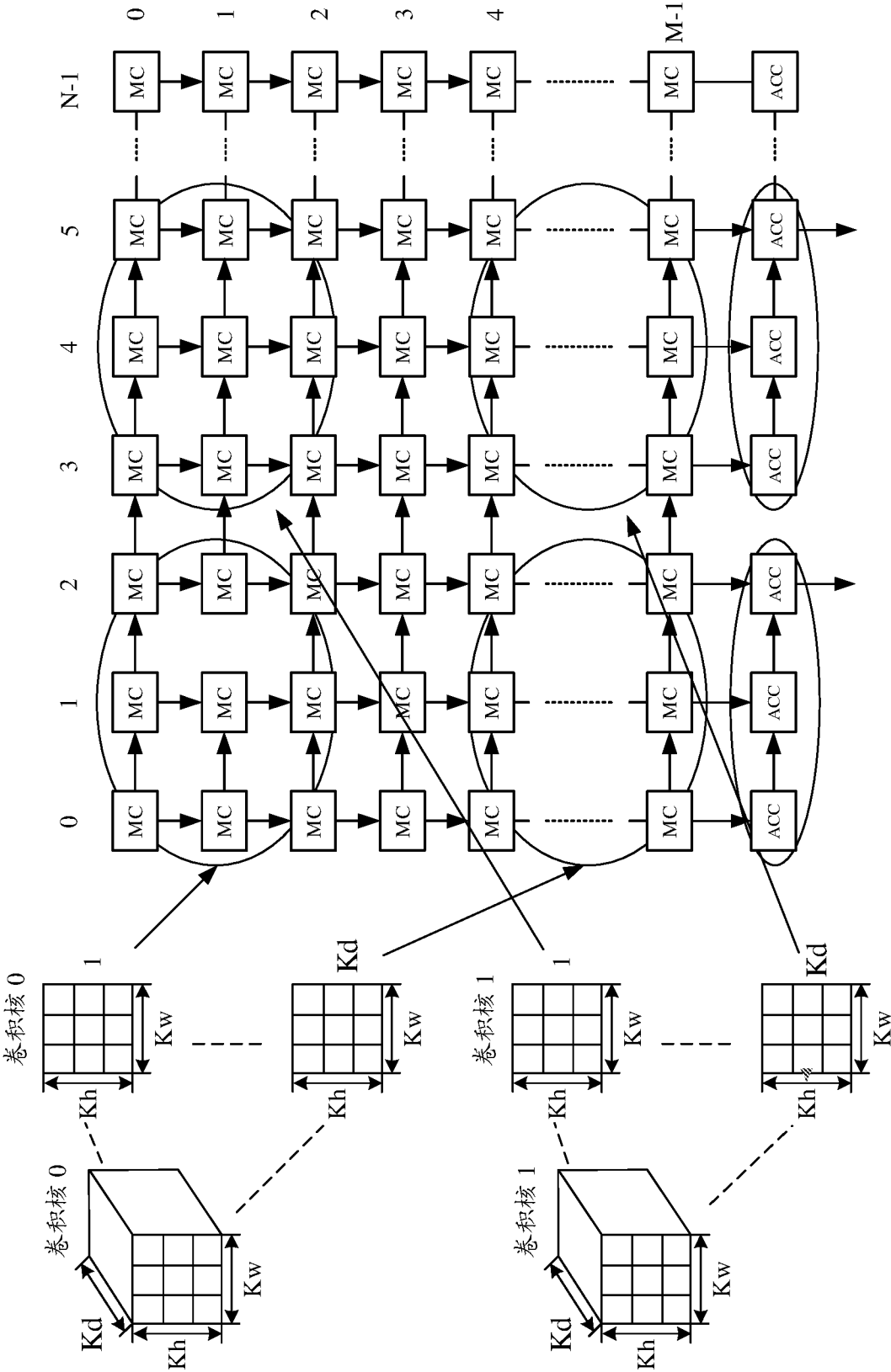


图6

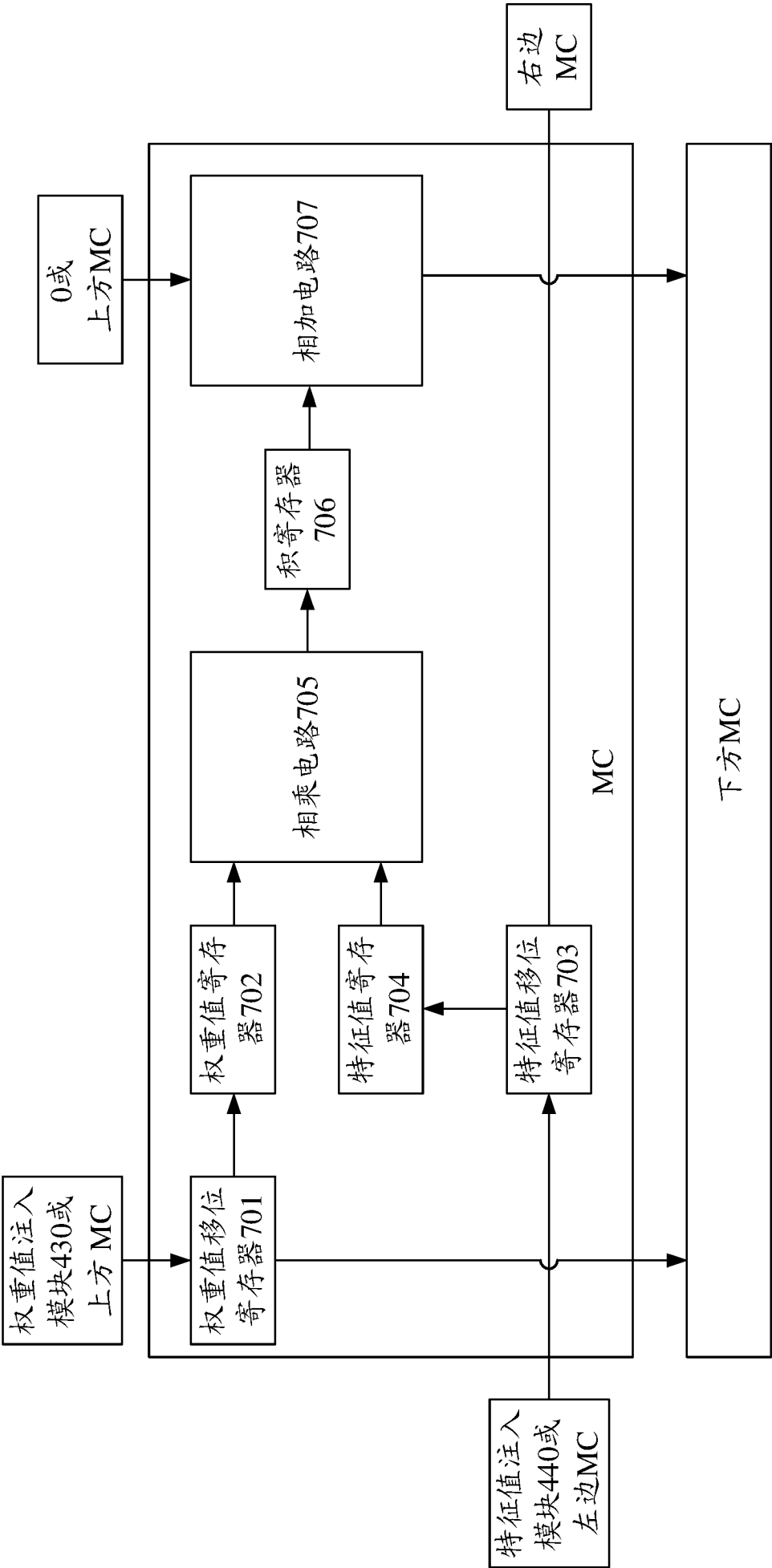


图7

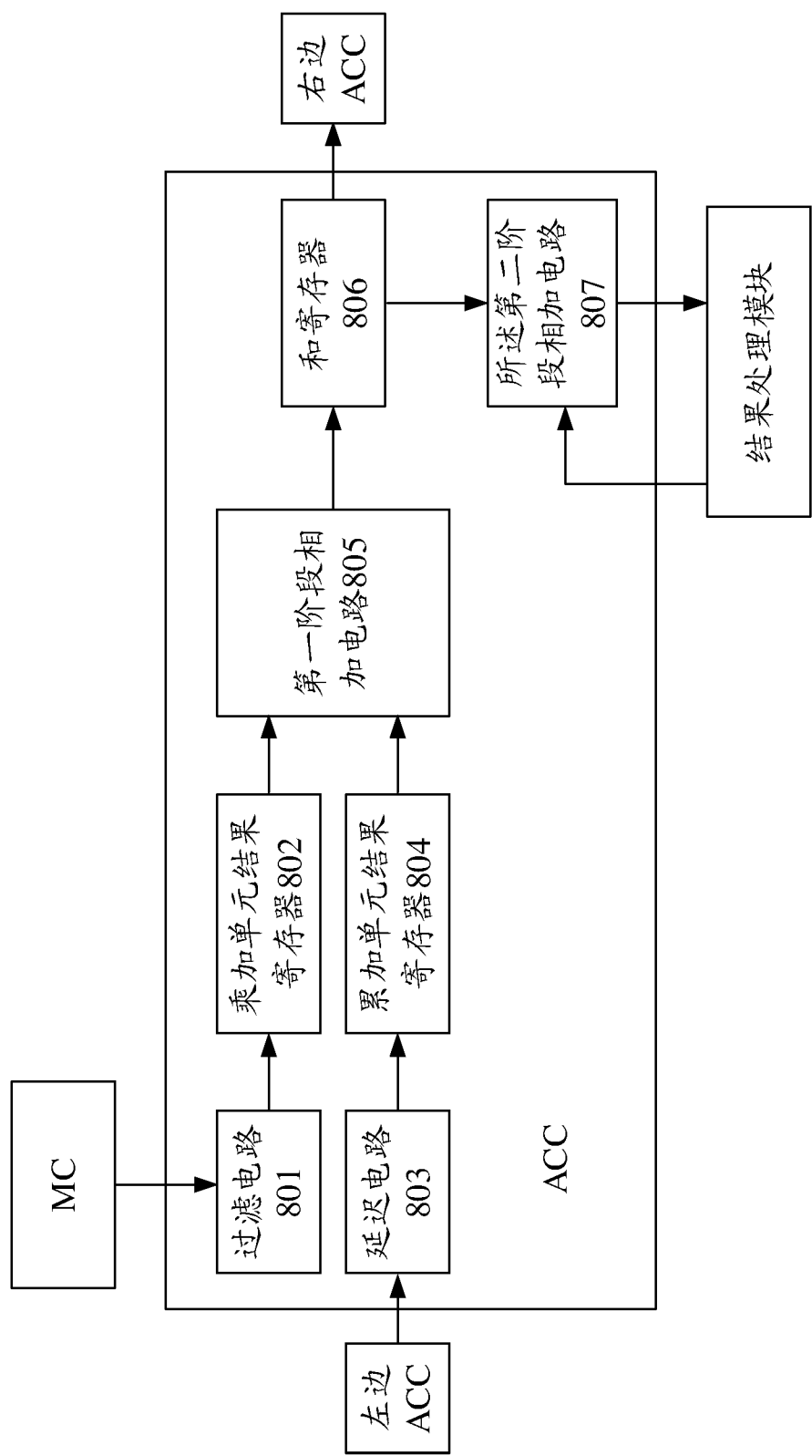


图 8

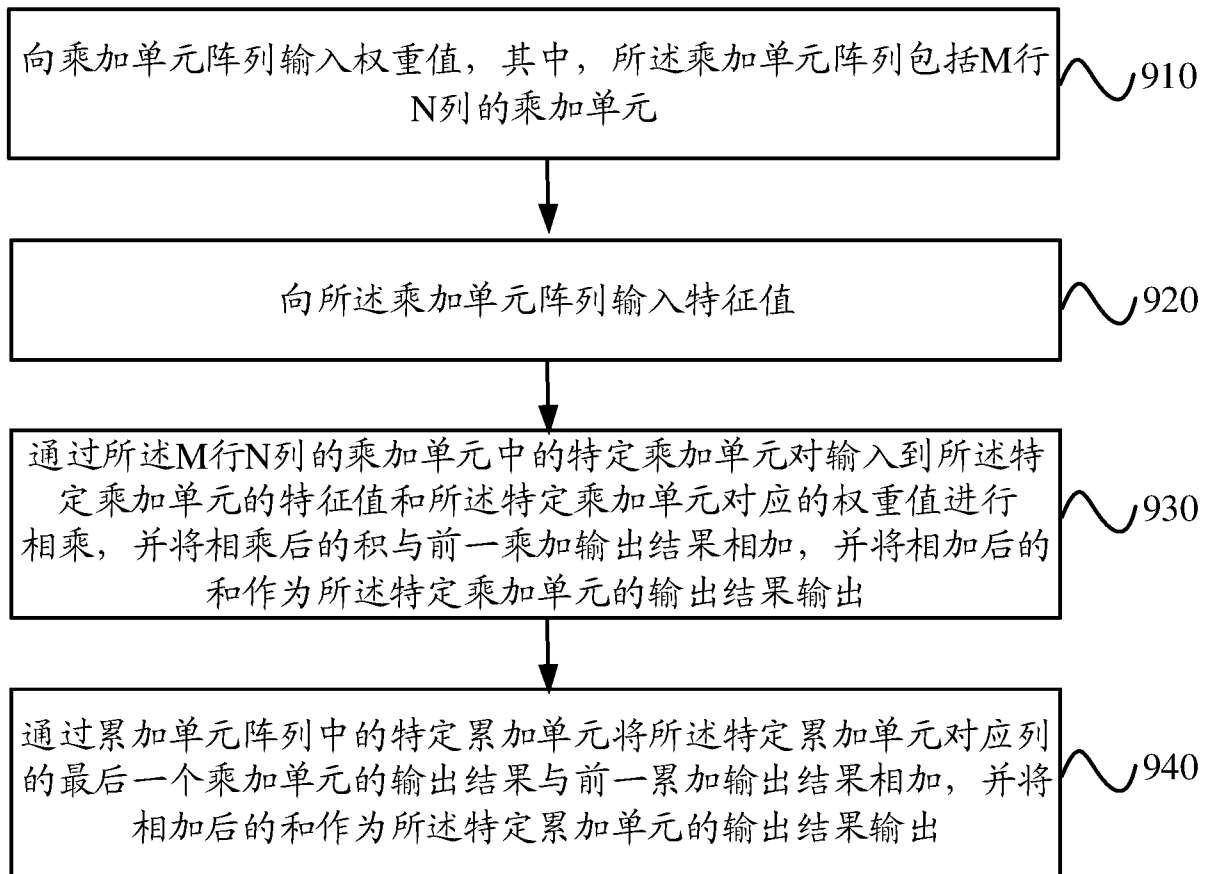
900

图9

## INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/074249

**A. CLASSIFICATION OF SUBJECT MATTER**

G06N 3/063(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

**B. FIELDS SEARCHED**

Minimum documentation searched (classification system followed by classification symbols)

G06N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, CNKI, WPI, EPODOC: 神经网络, 卷积, 乘加, 累加, 阵列, 矩阵, 过滤, 延迟, 膨胀值, neural network, convolution, CNN, multiply, add, accumulate, MAU, ACC, array, filter, dilation, delay

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

| Category* | Citation of document, with indication, where appropriate, of the relevant passages                                                                                   | Relevant to claim No. |
|-----------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------|
| Y         | CN 107918794 A (INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES) 17 April 2018 (2018-04-17)<br>description, paragraphs [0038]-[0089], and figures 2-6 | 1-11, 14-24, 27-29    |
| Y         | CN 109284821 A (HUAWEI TECHNOLOGIES CO., LTD.) 29 January 2019 (2019-01-29)<br>description, paragraph [0029], and figure 4                                           | 1-11, 14-24, 27-29    |
| A         | CN 108537330 A (INSTITUTE OF AUTOMATION, CHINESE ACADEMY OF SCIENCES) 14 September 2018 (2018-09-14)<br>entire document                                              | 1-29                  |
| A         | US 2016328838 A1 (APPLIED RESEARCH LLC.) 10 November 2016 (2016-11-10)<br>entire document                                                                            | 1-29                  |
| A         | CN 108520297 A (ZHOU, Jun) 11 September 2018 (2018-09-11)<br>entire document                                                                                         | 1-29                  |



Further documents are listed in the continuation of Box C.



See patent family annex.

\* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

18 October 2019

Date of mailing of the international search report

04 November 2019

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/  
CN)  
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing  
100088  
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

**INTERNATIONAL SEARCH REPORT**  
**Information on patent family members**

International application No.

**PCT/CN2019/074249**

| Patent document<br>cited in search report |            |    | Publication date<br>(day/month/year) | Patent family member(s) | Publication date<br>(day/month/year) |
|-------------------------------------------|------------|----|--------------------------------------|-------------------------|--------------------------------------|
| CN                                        | 107918794  | A  | 17 April 2018                        | None                    |                                      |
| CN                                        | 109284821  | A  | 29 January 2019                      | None                    |                                      |
| CN                                        | 108537330  | A  | 14 September 2018                    | None                    |                                      |
| US                                        | 2016328838 | A1 | 10 November 2016                     | None                    |                                      |
| CN                                        | 108520297  | A  | 11 September 2018                    | None                    |                                      |

## 国际检索报告

国际申请号

PCT/CN2019/074249

## A. 主题的分类

G06N 3/063 (2006.01) i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

## B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06N

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

CNPAT, CNKI, WPI, EP0D0C: 神经网络, 卷积, 乘加, 累加, 阵列, 矩阵, 过滤, 延迟, 膨胀值, neural network, convolution, CNN, multiply, add, accumulate, MAU, ACC, array, filter, dilation, delay

## C. 相关文件

| 类 型* | 引用文件, 必要时, 指明相关段落                                                             | 相关的权利要求            |
|------|-------------------------------------------------------------------------------|--------------------|
| Y    | CN 107918794 A (中国科学院计算技术研究所) 2018年 4月 17日 (2018 - 04 - 17)<br>说明书38-89段、图2-6 | 1-11, 14-24, 27-29 |
| Y    | CN 109284821 A (华为技术有限公司) 2019年 1月 29日 (2019 - 01 - 29)<br>说明书29段、图4          | 1-11, 14-24, 27-29 |
| A    | CN 108537330 A (中国科学院自动化研究所) 2018年 9月 14日 (2018 - 09 - 14)<br>全文              | 1-29               |
| A    | US 2016328838 A1 (APPLIED RESEARCH LLC.) 2016年 11月 10日 (2016 - 11 - 10)<br>全文 | 1-29               |
| A    | CN 108520297 A (周军) 2018年 9月 11日 (2018 - 09 - 11)<br>全文                       | 1-29               |

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

\* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&amp;” 同族专利的文件

国际检索实际完成的日期

2019年 10月 18日

国际检索报告邮寄日期

2019年 11月 4日

ISA/CN的名称和邮寄地址

中国国家知识产权局(ISA/CN)

中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

受权官员

董立波

电话号码 86-(10)-53961431

国际检索报告  
关于同族专利的信息

国际申请号

PCT/CN2019/074249

| 检索报告引用的专利文件 |            |    | 公布日<br>(年/月/日) | 同族专利 | 公布日<br>(年/月/日) |
|-------------|------------|----|----------------|------|----------------|
| CN          | 107918794  | A  | 2018年 4月 17日   | 无    |                |
| CN          | 109284821  | A  | 2019年 1月 29日   | 无    |                |
| CN          | 108537330  | A  | 2018年 9月 14日   | 无    |                |
| US          | 2016328838 | A1 | 2016年 11月 10日  | 无    |                |
| CN          | 108520297  | A  | 2018年 9月 11日   | 无    |                |