

(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 2 月 20 日 (20.02.2020)



(10) 国际公布号
WO 2020/034079 A1

- (51) 国际专利分类号:
G06N 3/06 (2006.01)
- (21) 国际申请号: PCT/CN2018/100335
- (22) 国际申请日: 2018 年 8 月 14 日 (14.08.2018)
- (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: 高明明 (GAO, Mingming); 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,

(54) Title: SYSTOLIC ARRAY-BASED NEURAL NETWORK PROCESSING DEVICE

(54) 发明名称: 基于脉动阵列的神经网络处理装置

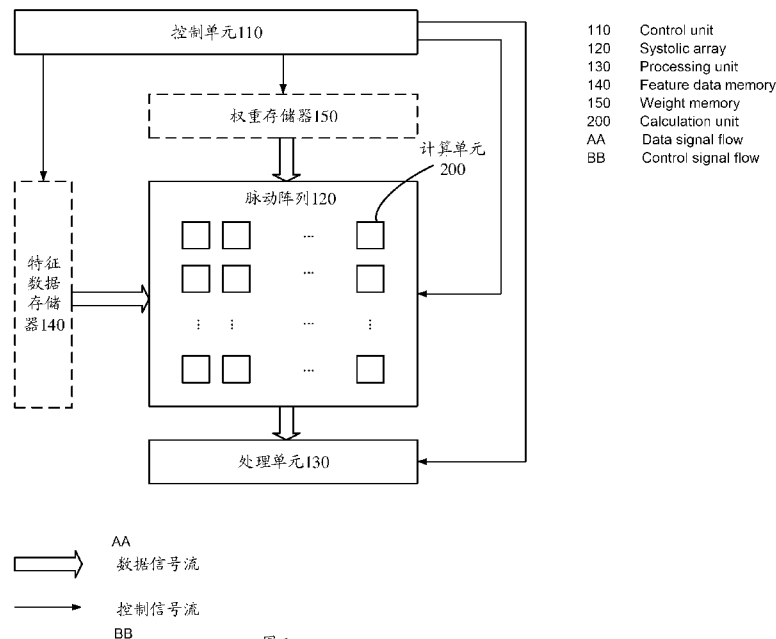


图 1

(57) Abstract: Provided is a neural network processing device. The device comprises: a systolic array; a control unit, used to transmit a control signal indicating a convolution operation, an average pooling operation, or a maximization operation, wherein each calculation unit in the systolic array is used to receive the control signal, input feature data, a weight, and a first intermediate processing result from an upper level calculation unit along a second dimension of the systolic array, performing a related operation on the input feature data, the weight, and the first intermediate processing result according to the control signal, so as to acquire a second intermediate

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, 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)。

根据细则4.17的声明:

- 关于申请人有权申请并被授予专利 (细则4.17(ii))

本国际公布:

- 包括国际检索报告 (条约第21条(3))。

processing result, transmitting the second intermediate processing result to a lower level calculation unit along the second dimension, and transmitting the input feature data to a lower level calculation unit along the first dimension; and a processing unit, used to receive the control signal and the second intermediate processing result outputted by a calculation unit of an output end of the systolic array, and perform a related operation on the second intermediate processing result according to the control signal, so as to acquire a neural network output result. The device reduces hardware overheads.

(57) 摘要: 提供一种神经网络处理装置, 该装置包括: 脉动阵列; 控制单元, 用于发送指示卷积操作、平均池化操作或最大化操作的控制信号; 脉动阵列中各个计算单元用于, 接收控制信号、输入特征数据、权重以及来自沿脉动阵列的第二维度的上一级计算单元的第一中间处理结果; 根据控制信号, 对输入特征数据、权重以及第一中间处理结果进行相关操作, 获得第二中间处理结果; 向沿着第二维度的下一级计算单元发送第二中间处理结果; 向沿着第一维度的下一级计算单元发送输入特征数据; 处理单元, 用于接收控制信号以及脉动阵列输出端的计算单元输出的第二中间处理结果, 根据控制信号对第二中间处理结果进行相关操作, 获得神经网络输出结果。可以降低硬件开销。

基于脉动阵列的神经网络处理装置

版权申明

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

本申请涉及神经网络领域，并且更为具体地，涉及一种基于脉动阵列的神经网络处理装置。

10

背景技术

随着神经网络复杂度的不断提高，神经网络技术在实际应用过程中存在占用资源多、运算速度慢、能量消耗大等问题。采用硬件加速器替代传统软件计算的方法成为提高神经网络计算效率的行之有效的方式，例如利用通用图形处理器、专用处理器芯片和现场可编程逻辑阵列实现的神经网络处理器。

当前技术，针对神经网络中的卷积操作设计一套独立的硬件装置，针对池化操作设计一套独立的硬件装置，或者，分别针对平均池化操作与最大池化操作，设计一套独立的硬件装置。

现有方案，在一定程度上增加了神经网络的硬件开销与功耗。

20

发明内容

本申请提供一种基于脉动阵列的神经网络处理装置，可以有效降低神经网络的硬件开销与功耗。

第一方面，提供一种神经网络处理装置，该装置包括：脉动阵列，包括多个计算单元；控制单元，用于发送控制信号，该控制信号指示如下神经网络操作中的任一种：卷积操作、平均池化操作以及最大化操作；该脉动阵列中的各个计算单元用于，接收该控制信号、输入特征数据、权重以及来自沿该脉动阵列的第二维度的上一级计算单元的第一中间处理结果；根据该控制信号，对该输入特征数据、该权重以及该第一中间处理结果进行神经网络相关操作，获得第二中间处理结果；向沿着该脉动阵列的该第二维度的下一级计算单元发送该第二中间处理结果；向沿着该脉动阵列的第一维度的下一级

计算单元发送该输入特征数据；处理单元，与该脉动阵列输出端的计算单元连接，该处理单元用于接收该控制信号以及该脉动阵列输出端的计算单元输出的第二中间处理结果，根据该控制信号对该第二中间处理结果进行神经网络相关操作，获得神经网络输出结果。

- 5 第二方面，提供一种池化处理装置，该装置包括：脉动阵列，包括多个计算单元；控制单元，用于发送控制信号，该控制信号指示平均池化操作或最大化操作；该脉动阵列中的各个计算单元用于，接收该控制信号、输入特征数据以及来自沿该脉动阵列的第二维度的上一级计算单元的第一中间处理结果；根据该控制信号，对该输入特征数据以及该第一中间处理结果进行
- 10 池化相关操作，获得第二中间处理结果；向沿着该脉动阵列的该第二维度的下一级计算单元发送该第二中间处理结果；向沿着该脉动阵列的第一维度的下一级计算单元发送该输入特征数据；处理单元，与该脉动阵列输出端的计算单元连接，该处理单元用于接收该控制信号以及该脉动阵列输出端的计算单元输出的第二中间处理结果，根据该控制信号对该第二中间处理结果进行
- 15 池化相关操作，获得池化输出结果。

第三方面，提供一种芯片，该芯片上集成第一方面提供的装置或集成第二方面提供的装置。

第四方面，提供一种控制方法，该控制方法由第一方面提供的装置中的控制单元执行，该方法包括第一方面中的控制单元执行的相关流程。

- 20 第五方面，提供一种控制装置，该控制装置包括存储器和处理器，该存储器用于存储指令，该处理器用于执行该存储器存储的指令，并且对该存储器中存储的指令的执行使得该处理器执行第四方面提供的方法。

第六方面，提供一种计算机可读存储介质，其上存储有计算机程序，该计算机程序被计算机执行时使得该计算机实现第四方面提供的方法。

- 25 第七方面，提供一种包含指令的计算机程序产品，该指令被计算机执行时使得该计算机实现第四方面提供的方法。

附图说明

- 图 1 为根据本申请实施例的基于脉动阵列的神经网络处理装置的示意性
- 30 框图。

图 2 为采用本申请实施例提供的基于脉动阵列的神经网络处理装置实现

神经网络操作的示意性流程图。

图 3 至图 6 为本申请实施例中的计算单元的内部结构的示意图。

图 7 至图 14 为采用本申请实施例提供的神经网络处理装置实现卷积操作或平均池化操作的时序示意图。

5 图 15 至图 21 为采用本申请实施例提供的神经网络处理装置实现最大池化操作的时序示意图。

图 22 为采用本申请实施例提供的神经网络处理装置实现针对多个卷积核的卷积操作的示意图。

10 图 23 为根据本申请另一实施例提供的基于脉动阵列的池化处理装置的示意性框图。

具体实施方式

下面将结合附图，对本申请实施例中的技术方案进行描述。

15 除非另有定义，本文所使用的所有的技术和科学术语与属于本申请的技术领域的技术人员通常理解的含义相同。本文中在本申请的说明书中所使用的术语只是为了描述具体的实施例的目的，不是旨在于限制本申请。

图 1 为根据本申请实施例的神经网络处理装置 100 的示意性框图。该装置 100 包括控制单元 110、脉动阵列 120、处理单元 130。

20 控制单元 110，用于向脉动阵列 120 和处理单元 130 发送控制信号，控制信号指示如下神经网络操作中的任一种：卷积操作、平均池化操作以及最大化操作。

脉动阵列 120，包括多个计算单元 200。脉动阵列 120 中的每个计算单元 200 的内部结构相同，下文将详细进行描述。

25 脉动阵列 120 中的各个计算单元 200 用于，接收控制信号、输入特征数据、权重以及来自沿脉动阵列的第二维度的上一级计算单元的第一中间处理结果；根据控制信号，对输入特征数据、权重以及第一中间处理结果进行神经网络相关操作，获得第二中间处理结果；向沿着脉动阵列的第二维度的下一级计算单元发送第二中间处理结果；向沿着脉动阵列的第一维度的下一级计算单元发送输入特征数据。

30 当控制信号指示卷积操作时，计算单元 200 对输入特征数据、权重以及第一中间处理结果进行的神经网络相关操作包括乘法运算和加法运算。当控

制信号指示平均池化操作时，计算单元 200 对输入特征数据、权重以及第一中间处理结果进行的神经网络相关操作包括加法运算。当控制信号指示最大池化操作时，计算单元 200 对输入特征数据、权重以及第一中间处理结果进行的神经网络相关操作包括比较运算。

- 5 第一维度对应于脉动阵列的行，第二维度对应于脉动阵列的列。或者第一维度对应于脉动阵列的列，第二维度对应于脉动阵列的行。

为了便于理解与描述，下文某些实施例中以第一维度对应于脉动阵列的行，第二维度对应于脉动阵列的列为例进行描述。例如，在图 1 所示实施例中，第一维度对应于脉动阵列 120 的行，第二维度对应于脉动阵列 140 的列。

- 10 下文某些实施例中也会出现“从左到右”与“从上到下”的描述，其中“从左到右”表示沿着脉动阵列的行从左到右，“从上到下”表示沿着脉动阵列的列从上到下。

如图 1 所示，对于脉动阵列 120 中沿着第一维度上的输入端的计算单元来说，其可以从脉动阵列 120 之外的特征数据存储器 140 中接收输入特征值。

- 15 其余计算单元均是从沿着第一维度的上一级计算单元接收输入特征值。

如图 1 所示，脉动阵列 120 中各个计算单元均可以从脉动阵列 120 之外的权重存储器 150 中接收权重。

- 20 可选地，由主处理器确定需要载入脉动阵列 120 中的输入特征数据和权重，然后将该输入特征数据从特征数据存储器 140 载入到脉动阵列 120 中，将该权重从权重存储器 150 载入到脉动阵列 120 中。其中，该主处理器可以为控制单元 110，也可以为神经网络系统中的了解脉动阵列的计算规律的其它处理器。

可选地，本申请提供的神经网络处理装置 100 包括特征数据存储器 140 和权重存储器 150。

- 25 可选地，本申请提供的神经网络处理装置 100 不包括特征数据存储器 140 和权重存储器 150，但是包括用于与特征数据存储器 140 和权重存储器 150 进行通信的通信接口。

- 30 在本实施例中，神经网络处理装置 100 还可以包括一个缓存单元，该缓存单元用于对从通信接口传递过来的数据进行缓存。也就是说，特征数据存储器 140 和权重存储器 150 中的数据通过该缓存单元，进入脉动阵列 120 中。应理解，利用缓存单元缓存从通信接口传递过来的数据，有助于保证进入脉

动阵列 120 的数据的时序的正确性。

处理单元 130，与脉动阵列 120 输出端的计算单元 200 连接，处理单元 130 用于接收控制信号以及脉动阵列输出端的计算单元输出的第二中间处理结果，根据控制信号对第二中间处理结果进行神经网络相关操作，获得神经网络输出结果。

本文中描述的某个单元与某个单元连接，包括但不限于有线通信连接和无线通信连接。

脉动阵列 120 输出端的计算单元指的是，脉动阵列 120 中最终输出中间处理结果的计算单元，例如，在图 1 中，指的是，脉动阵列 120 的最后一行的计算单元。

例如，脉动阵列 120 输出端的计算单元的发送接口与处理单元 130 的接收接口通信连接。

当控制信号指示卷积操作时，处理单元 130 对第二中间处理结果进行累加运算。当控制信号指示平均池化操作时，处理单元 130 对第二中间处理结果进行求平均运算。当控制信号指示最大池化操作时，处理单元 130 对第二中间处理结果进行比较运算。

上述可知，控制单元 110 用于控制脉动阵列 120 和处理单元 130 进行神经网络相关操作的类型，即卷积操作、平均池化操作或最大池化操作，从而使得本申请提供的装置 100 既可以实现卷积操作，又可以实现池化操作。

因此，本申请提供的方案，通过使得卷积运算和池化运算复用一套硬件装置，从而可以提高硬件利用率，节省成本。

需要说明的是，在本申请中，在脉动阵列中脉动（在脉动阵列的各个计算单元之间转发）的元素包括输入特征数据与计算单元计算得到的中间处理结果，不包括权重。换言之，对于一次神经网络操作，权重被载入计算单元后，不再被转发。

应理解，本文中提及的第一中间处理结果与第二中间处理结果仅为了区分同一个计算单元接收与发送的中间处理结果，还可以采用其它描述来区分同一个计算单元接收与发送的中间处理结果，本申请对此不做限定。

还应理解，当一个计算单元 X 为脉动阵列沿着第一维度的输入端的计算单元时，该计算单元 X 不存在沿着第一维度的前一级计算单元，即该计算单元 X 从外部存储器中获取输入特征数据。当一个计算单元 Y 为脉动阵列沿

着第二维度的输入端的计算单元时,该计算单元 Y 不存在沿着第二维度的前一级计算单元,即该计算单元 Y 的用于接收上一级计算单元的第一中间处理结果的端口的输入为空或者为零。当一个计算单元 Z 为脉动阵列沿着第二维度的输出端的计算单元时,该计算单元 Z 不存在沿着第二维度的后一级计算单元,该计算单元 Z 与处理单元 130 连接,用于将获得的第二中间处理结果发送到处理单元 130。

应理解,在采用本申请提供的装置 100 实现卷积操作时,载入到脉动阵列 120 中的权重就是卷积核。

可选地,在控制信号指示卷积操作时,权重在脉动阵列 120 中的分布方式与权重在卷积核中的分布方式一致或呈镜像。

例如,卷积核对应的二维矩阵为 $\begin{bmatrix} W_{11} & W_{12} & W_{13} \\ W_{21} & W_{22} & W_{23} \\ W_{31} & W_{32} & W_{33} \end{bmatrix}$,则载入到脉动阵列 120

中的权重在脉动阵列 120 中的分布方式可以为 $\begin{matrix} W_{11} & W_{12} & W_{13} \\ W_{21} & W_{22} & W_{23} \\ W_{31} & W_{32} & W_{33} \end{matrix}$,也可以为

$\begin{matrix} W_{13} & W_{12} & W_{11} \\ W_{23} & W_{22} & W_{21} \\ W_{33} & W_{32} & W_{31} \end{matrix}$ 。

由主处理器(例如控制单元 110 或神经网络系统中的其他处理器)从权重存储器中读取相应的权重载入到脉动阵列 120 中,其中,权重的寻址以及权重在脉动阵列 120 中的分布方式均可以根据实际需求来确定,本申请对此不做限定,只要能够实现卷积操作或池化操作即可。

池化操作仅针对输入特征数据,并不涉及权重,因此,为了能够使计算单元 200 可以兼容卷积运算和池化运算,在控制信号指示最大池化操作或平均池化操作时,将权重的值配置为 1。

可选地,在一些实施例,装置 100 还包括如下单元。

特征数据存储器(如图 1 中所示的特征数据存储器 140),用于存储输入特征数据。

第一权重存储器(如图 1 中所示的权重存储器 150)和第二权重存储器(图 1 中未示出)。当控制信号指示卷积操作时,第一权重存储器用于存储当前卷积操作的卷积核,第二权重存储器用于预先存储下一次卷积操作的卷

积核。

控制单元 110 用于将第一权重存储器中的权重载入脉动阵列 120 中的计算单元，即载入沿着脉动阵列 120 的第一维度的输入端的计算单元中。

5 本实施例提供的装置，在执行卷积操作时，通过预先存储当前卷积操作的下一次卷积操作的卷积核，从而在利用脉动阵列完成当前卷积操作的相关运算后，可以立即开始下一次卷积操作的运算，从而可以提高卷积操作的效率。

作为示例，如图 2 所示，采用如图 1 所示的基于脉动阵列的装置 100 处理神经网络的过程如下。

10 S210，控制单元 110 确定神经网络操作类型，神经网络操作类型为下列操作中的任一种：卷积操作、平均池化操作以及最大池化操作。

S220，控制单元 110 根据神经网络操作类型，对存储单元（如图 1 中所示的特征存储器与权重存储器）寻址，读取并解析下一步需要执行的指令。

15 S230，控制单元 110 根据解析指令得到存储地址，从存储单元中获取输入数据（输入特征数据和权重），将输入特征数据和权重分别载入至脉动阵列中。

S240，控制单元 110 向脉动阵列 120 中的各个计算单元 200 发送控制信号，该控制信号指示神经网络操作类型。

20 S250，脉动阵列 120 中的各个计算单元 200 基于控制信号，对接收的输入特征数据与权重执行相关运算，获得第二中间处理结果，并向沿着第二维度的下一级计算单元发送该第二中间处理结果。下文将详细描述步骤 S250。

S260，控制单元 120 向处理单元 130 发送控制信号，该控制信号指示神经网络操作类型。

25 S270，处理单元 130 从脉动阵列 120 输出端的计算单元 200 获取第二中间处理结果，并基于控制信号，对该第二中间处理结果执行相关运算，获得神经网络的输出结果。下文将详细描述步骤 S270。

可选地，处理单元 130 输出的神经网络的输出结果可以被存储到神经网络系统的输出存储单元中。

30 以脉动阵列中的某一个计算单元 200 为例，上述步骤 S250 包括如下内容。

当控制信号指示卷积操作或平均池化操作时，计算单元 200 对输入特征

数据与权重进行相乘，获得乘积，对该乘积与第一中间处理结果进行累加，获得累加结果，该累加结果为第二中间处理结果。

当控制信号指示最大池化操作时，计算单元 200 对输入特征数据与第一中间处理结果进行比较，获得二者中的最大值，该最大值为第二中间处理结果。或者，计算单元 200 对输入特征数据与权重进行相乘，获得乘积，对该乘积与第一中间处理结果进行比较，获得二者中的最大值，该最大值为第二中间处理结果。

需要说明的是，当控制信号指示池化操作（平均池化或最大池化）时，载入脉动阵列中的权重的值均为 1。

应理解，当执行步骤 S250 的计算单元为脉动阵列中沿着第二维度输出端的计算单元，则向处理单元 130 发送第二中间处理结果。

以处理单元 130 接收到第二中间处理结果为某一个卷积核对应的卷积操作的中间结果为例，或者，以处理单元 130 接收到第二中间处理结果为某一个池化窗对应的池化操作的中间结果为例，步骤 S270 具体包括如下内容。

当控制信号指示卷积操作时，处理单元 130 对接收的多个第二中间处理结果进行累加，获得一个卷积核的计算结果。

当控制信号指示平均池化操作时，处理单元 130 基于池化窗的大小($k*k$)，对接收的多个第二中间处理结果进行求平均，获得一个池化窗的平均池化计算结果。

当控制信号指示最大池化操作时，处理单元 130 对接收的多个第二中间处理结果进行比较，获得其中的最大值，则获得了一个池化窗的最大池化计算结果。

在本申请实施例中，脉动阵列中的每个计算单元采用相同的内部结构。本申请中，计算单元的内部结构可以有多种实现方式。

图 3 和图 4 是计算单元 200 的一种内部结构的示意图。图 3 和图 4 分别是计算单元 200 基于同一种内部结构的两种工作模式。

图 5 和图 6 是计算单元 200 的另一种内部结构的示意图。图 5 和图 6 分别是计算单元 200 基于同一种内部结构的两种工作模式。

作为一种实现方式，计算单元 200 的内部结构如图 3（或图 4）所示。

特征寄存器 201 用于存储输入特征数据，还用于向沿着脉动阵列的第一维度的后一级计算单元转发输入特征数据。

权重寄存器 202 用于存储权重。

乘法器 203 与特征寄存器 201 和权重寄存器 202 连接，乘法器 203 用于输出权重与输入特征数据的乘积。

加法器 204 与乘法器 203 连接。

5 比较器 205，与乘法器 203 连接。

基于这种内部结果，计算单元 200 具有两种工作模式。

具体地，计算单元 200 接收控制信号，当解析获得控制信号指示卷积操作或平均池化操作时，开启工作模式一，如图 3 中所示。加法器 204 用于，接收乘法器 203 输出的乘积与来自沿着第二维度的前一级计算单元的第一中间处理结果，对乘积与第一中间处理结果进行累加，获得第二中间处理结果，
10 向沿着第二维度的后一级计算单元发送第二中间处理结果。在工作模式一，比较器 205 所在的电路是无效的（如图 3 中虚线标注的电路）。

计算单元 200 接收控制信号，当解析获得控制信号指示最大池化操作时，开启工作模式二，如图 4 所示。比较器 205 用于接收乘法器 203 输出的乘积与来自沿着第二维度的前一级计算单元的第一中间处理结果，并输出乘积与第一中间结果中的最大值，向沿着第二维度的后一级计算单元发送第二中间处理结果，第二中间处理结果为最大值。在工作模式二，加法器 204 所在的电路是无效的（如图 4 中虚线标注的电路）。

计算单元 200 还包括一个控制子单元，该控制子单元用根据控制信号切换计算单元 200 的工作模式。
20

作为另一种实现方式，计算单元 200 的内部结构如图 5（或图 6）所示。

特征寄存器 401 用于存储输入特征数据，还用于向沿着脉动阵列的第一维度的后一级计算单元转发输入特征数据。

权重寄存器 402 用于存储权重。

25 乘法器 403 与特征寄存器 401 和权重寄存器 402 连接，乘法器 403 用于输出权重与输入特征数据的乘积。

加法器 404 与乘法器 403 连接。

选择器 405 与乘法器 403 和加法器 404 连接。

基于这种内部结果，计算单元 200 具有两种工作模式。

30 具体地，计算单元 200 接收控制信号，当解析获得控制信号指示卷积操作或平均池化操作时，开启工作模式一，如图 5 中所示。加法器 404 用于，

接收乘法器 403 输出的乘积与来自沿着第二维度的前一级计算单元的第一中间处理结果,对乘积与第一中间处理结果进行累加,获得第二中间处理结果,向沿着第二维度的后一级计算单元发送第二中间处理结果。在工作模式一,选择器 405 所在的电路是无效的(如图 5 中虚线标注的电路)。

- 5 计算单元 200 接收控制信号,当解析获得控制信号指示最大池化操作时,开启工作模式二,如图 6 所示。加法器 404 用于,接收乘法器 403 输出的乘积与来自沿着第二维度的前一级计算单元的第一中间处理结果,输出乘积与第一中间处理结果的相减结果;选择器 405 用于接收乘法器 403 输出的乘积、加法器 404 输出的相减结果、以及第一中间处理结果,根据相减结果从乘积与第一中间处理结果中选择出最大值,并向沿着第二维度的后一级计算单元发送第二中间处理结果,第二中间处理结果为最大值。在工作模式二,加法器的 404 所在的某些电路是无效的(如图 6 中虚线标注的电路)。

例如,计算单元 200 还包括一个控制子单元,该控制子单元用根据控制信号切换计算单元 200 的工作模式。

- 15 应理解,本文中提及的乘法器可以采用乘法电路实现,加法器可以采用加法电路实现,比较器可以采用比较电路实现,选择器可以采用比较电路实现。

- 20 应理解,图 3 至图 6 仅为示例而非限定。计算单元 200 的内部结果还可以采用其他实现方式,本申请对比不做限定,只要能够使得计算单元 200 根据控制信号所指示的神经网络操作类型切换工作模式即可。

处理单元 130 根据控制信号所指示的神经网络操作类型,可以在多种工作模式之间切换。

在控制信号指示卷积操作时,处理单元 130 用于对接收的第二中间处理结果进行累加,获得卷积操作的输出结果。

- 25 在控制信号指示平均池化操作时,处理单元 130 用于基于池化窗的大小,对接收的第二中间处理结果进行求平均,获得平均池化操作的输出结果。

在控制信号指示最大池化操作时,处理单元 130 用于对接收的第二中间处理结果进行比较,获得最大池化操作的输出结果。

- 30 作为一个示例,处理单元 130 的内部结果包括:累加器、除法器(或乘法器)、比较器。

在控制信号指示卷积操作时,处理单元 130 内部的累加器工作。具体地,

该累加器用于输出多个计算单元的第二中间处理结果的累加结果，获得卷积操作的输出结果。

在控制信号指示平均池化操作时，处理单元 130 内部的累加器与除法器工作。具体地，该累加器用于输出多个计算单元的第二中间处理结果的累加结果，除法器用于基于池化窗口的大小，对该累加结果求平均，例如，池化窗口的大小为 $k*k$ ，则除法器将该累加结果除以 $k*k$ ，获得平均池化操作的输出结果。应理解，求平均也可以采用乘法器来实现，例如，池化窗口的大小为 $k*k$ ，则乘法器将该累加结果乘以 $1/(k*k)$ ，获得平均池化操作的输出结果。

在控制信号指示最大池化操作时，处理单元 130 内部的比较器工作。具体地，该比较器用于输出多个计算单元的第二中间处理结果中的最大值，获得最大池化操作的输出结果。

因此，本申请提供的方案，通过使得卷积运算和池化运算复用一套硬件装置，从而可以提高硬件利用率，节省成本。

为了更好地理解本申请提供的方案，下文结合图 7 至图 21，以 $3*3$ 的脉动阵列、 $3*3$ 的权重矩阵、 $3*3$ 的输入特征矩阵为例，描述本申请提供的基于脉动阵列的神经网络处理装置执行卷积操作、平均池化操作以及最大池化操作的过程。

如图 7 所示，假设图 1 中所示的脉动阵列 120 由计算单元 C00、C01、C02、C10、C11、C12、C20、C21 与 C22 组成，处理单元 130 与计算单元 C20、C21 与 C22 连接，用于根据其输出的中间计算结果获得最终的计算结果。各个计算单元的内部结构可以为图 3 与图 4 所示，或者为图 5 或图 6 所示。

本例中，以 $3*3$ 的权重矩阵为 $\begin{bmatrix} W_{11} & W_{12} & W_{13} \\ W_{21} & W_{22} & W_{23} \\ W_{31} & W_{32} & W_{33} \end{bmatrix}$ 与 $3*3$ 的输入特征矩阵为

$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$ 为例进行描述。

第一种情形，控制信号指示卷积操作。

这种情形下，3*3 的权重矩阵 $\begin{bmatrix} W_{11} & W_{12} & W_{13} \\ W_{21} & W_{22} & W_{23} \\ W_{31} & W_{32} & W_{33} \end{bmatrix}$ 为卷积核。

对于数据块 $\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$ 和权重 $\begin{bmatrix} W_{11} & W_{12} & W_{13} \\ W_{21} & W_{22} & W_{23} \\ W_{31} & W_{32} & W_{33} \end{bmatrix}$ 的卷积操作，其计算结

果应为：

$$a_{11}*W_{11}+a_{12}*W_{12}+a_{13}*W_{13}+a_{21}*W_{21}+a_{22}*W_{22}+a_{23}*W_{23}+a_{31}*W_{31}+a_{32}*W_{32}+a_{33}*W_{33}。$$

如图 7 至图 14 所示，采用本申请的神经网络处理装置，执行输入特征数据与权重的载入和卷积计算的过程如下。

参见图 7 所示，预先将权重 $\begin{bmatrix} W_{11} & W_{12} & W_{13} \\ W_{21} & W_{22} & W_{23} \\ W_{31} & W_{32} & W_{33} \end{bmatrix}$ 加载到脉动阵列中。

下面实施例中描述的从左到右流动对应于上述实施例中描述的沿着脉动阵列的第一维度（行）流动。下面实施例中描述的从上到下流动对应于上述实施例中描述的沿着脉动阵列的第二维度（列）流动。

在第一个周期（T1），参见图 8 所示，输入特征数据 a_{11} 进入计算单元 C00，其中，输入特征数据 a_{11} 从计算单元 C00 的左边载入并从左向右流动。在第一个周期结束时，计算单元 C00 的计算结果为 $a_{11}*W_{11}$ 。计算单元 C00 的计算结果 $a_{11}*W_{11}$ 从上向下流动。

在第二个周期（T2），参见图 9 所示，输入特征数据 a_{11} 向右流动进入计算单元 C01，计算结果 $a_{11}*W_{11}$ 向下流动进入计算单元 C10；同时输入特征数据 a_{21} 载入至计算单元 C00，输入特征数据 a_{21} 载入至计算单元 C10。在第二个周期结束时，计算单元 C00 的计算结果为 $a_{12}*W_{11}$ ，计算单元 C01 的计算结果为 $a_{11}*W_{12}$ ，计算单元 C10 的计算结果为 $a_{11}*W_{11}+a_{21}*W_{21}$ 。各个计算单元的计算结果从上向下流动。

在第三个周期（T3），参见图 10 所示，输入特征数据 a_{11} 向右流动进入

计算单元 C02, a12 向右流动进入计算单元 C01, a21 向右流动进入计算单元 C11, 计算结果 $a12*W11$ 向下流动进入计算单元 C10, 计算结果 $a12*W12$ 向下流动进入计算单元 C11, 计算结果 $a11*W11+a21*W21$ 向下流动进入计算单元 C20。同时, a13 载入计算单元 C00, a22 载入计算单元 C10, a31 载入计算单元 C20。在第三个周期结束时, 计算单元 C00 的计算结果为 $a13*W11$, 计算单元 C01 的计算结果为 $a12*W12$, 计算单元 C02 的计算结果为 $a11*W13$, 计算单元 C10 的计算结果为 $a12*W11+a22*W21$, 计算单元 C11 的计算结果为 $a11*W12+a21*W22$, 计算单元 C20 的计算结果为 $a11*W11+a21*W21+a31*W31$ 。各个计算单元的计算结果从上向下流动。

10 在第四个周期 (T4), 参见图 11 所示, 输入特征数据 a12 向右流动进入计算单元 C02, a13 向右流动进入计算单元 C01, a21 向右流动进入计算单元 C12, a22 向右流动进入计算单元 C11, a31 向右流动进入计算单元 C21。第三个周期结束后, 计算单元 C00、C01、C02、C10、C11 的计算结果向下流动进入各自的下一级计算单元中。同时, a23 载入计算单元 C10, a32 载入计算单元 C20。在第四个周期结束时, 计算单元 C01 的计算结果为 $a13*W12$, 计算单元 C02 的计算结果为 $a12*W13$, 计算单元 C10 的计算结果为 $a13*W11+a23*W21$, 计算单元 C11 的计算结果为 $a12*W12+a22*W22$, 计算单元 C12 的计算结果为 $a11*W13+a21*W23$, 计算单元 C20 的计算结果为 $a12*W11+a22*W21+a32*W31$, 计算单元 C21 的计算结果为 $a11*W12+a21*W22+a31*W32$, 各个计算单元的计算结果从上向下流动。

20 在第五个周期 (T5), 参见图 12 所示, a13 向右流动进入计算单元 C02, a22 向右流动进入计算单元 C12, a23 向右流动进入计算单元 C11, a31 向右流动进入计算单元 C22, a32 向右流动进入计算单元 C21。第四个周期结束后, 计算单元 C01、C02、C10、C11、C12 的计算结果向下流动进入各自的下一级计算单元中。同时, a33 载入计算单元 C20。在第五个周期结束时, 计算单元 C02 的计算结果为 $a13*W13$, 计算单元 C11 的计算结果为

$a_{13} * W_{12} + a_{23} * W_{22}$, 计算单元 C12 的计算结果为 $a_{12} * W_{13} + a_{22} * W_{23}$, 计算单元 C20 的计算结果为 $a_{13} * W_{11} + a_{23} * W_{21} + a_{33} * W_{31}$, 计算单元 C21 的计算结果为 $a_{12} * W_{12} + a_{22} * W_{22} + a_{32} * W_{32}$, 计算单元 C22 的计算结果为 $a_{11} * W_{13} + a_{21} * W_{23} + a_{31} * W_{33}$, 各个计算单元的计算结果从上向下流动。

5 在第六个周期(T6), 参见图 13 所示, a_{23} 向右流动进入计算单元 C12, a_{32} 向右流动进入计算单元 C22, a_{33} 向右流动进入计算单元 C21。第五个周期结束后计算单元 C02、C11、C12 的计算结果向下流动进入各自的下一级计算单元中。在第六个周期结束时, 计算单元 C12 的计算结果为 $a_{13} * W_{13} + a_{23} * W_{23}$, 计算单元 C21 的计算结果为 $a_{13} * W_{12} + a_{23} * W_{22} + a_{33} * W_{32}$, 计算单元 C22 的计算结果为 $a_{12} * W_{13} + a_{22} * W_{23} + a_{32} * W_{33}$, 各个计算单元的计算结果从上向下流动。

15 在第七个周期(T7), 参见图 14 所示, a_{33} 向右流动进入计算单元 C22, 计算结果 $a_{13} * W_{13} + a_{23} * W_{23}$ 向下流动进入计算单元 C22。在第七个周期结束时, 计算单元 C22 的计算结果为 $a_{13} * W_{13} + a_{23} * W_{23} + a_{33} * W_{33}$, 各个计算单元的计算结果从上向下流动。

可知, 第三个周期结束时计算单元 C20 的计算结果 $a_{11} * W_{11} + a_{21} * W_{21} + a_{31} * W_{31}$ 、第五个周期结束时计算单元 C21 的计算结果 $a_{12} * W_{12} + a_{22} * W_{22} + a_{32} * W_{32}$ 、以及第七个周期结束时计算单元 C22 的计算

结果 $a_{13} * W_{13} + a_{23} * W_{23} + a_{33} * W_{33}$ 的累加值为数据块 $\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$ 和权重

20 $\begin{bmatrix} W_{11} & W_{12} & W_{13} \\ W_{21} & W_{22} & W_{23} \\ W_{31} & W_{32} & W_{33} \end{bmatrix}$ 的卷积操作的计算结果。

处理单元 130 用于接收计算单元 C20、C21 与 C22 输出的中间计算结果, 并对第三个周期结束时计算单元 C20 的计算结果、第五个周期结束时计算单元 C21 的计算结果、以及第七个周期结束时计算单元 C22 的计算结果作累

加, 得到数据块 $\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$ 和权重 $\begin{bmatrix} W_{11} & W_{12} & W_{13} \\ W_{21} & W_{22} & W_{23} \\ W_{31} & W_{32} & W_{33} \end{bmatrix}$ 的卷积操作的计算结果。

第二种情形, 控制信号指示平均池化操作。

这种情形下, $3*3$ 的权重矩阵 $\begin{bmatrix} W_{11} & W_{12} & W_{13} \\ W_{21} & W_{22} & W_{23} \\ W_{31} & W_{32} & W_{33} \end{bmatrix}$ 可以表示池化窗的大小为

$3*3$, 这里所谓的权重矩阵 $\begin{bmatrix} W_{11} & W_{12} & W_{13} \\ W_{21} & W_{22} & W_{23} \\ W_{31} & W_{32} & W_{33} \end{bmatrix}$ 的值为 $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$ 。

5 对于数据块 $\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$ 的平均池化操作, 其计算结果应为:

$(a_{11}+a_{12}+a_{13}+a_{21}+a_{22}+a_{23}+a_{31}+a_{32}+a_{33})/9$ 。

采用本申请的神经网络处理装置, 执行输入特征数据与权重的载入和平均池化计算的过程与上述的执行卷积运算的过程类似, 不同的是, 权重矩阵

10 $\begin{bmatrix} W_{11} & W_{12} & W_{13} \\ W_{21} & W_{22} & W_{23} \\ W_{31} & W_{32} & W_{33} \end{bmatrix}$ 的值为 $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$, 以及处理单元 130 基于 $3*3$ 的池化窗, 对第

三个周期结束时计算单元 C20 的计算结果、第五个周期结束时计算单元 C21 的计算结果、以及第七个周期结束时计算单元 C22 的计算结果求平均, 得到

数据块 $\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$ 的平均池化操作的计算结果

$(a_{11}+a_{12}+a_{13}+a_{21}+a_{22}+a_{23}+a_{31}+a_{32}+a_{33})/9$ 。

15 第三种情形, 控制信号指示最大池化操作。

这种情形下, $3*3$ 的权重矩阵 $\begin{bmatrix} W_{11} & W_{12} & W_{13} \\ W_{21} & W_{22} & W_{23} \\ W_{31} & W_{32} & W_{33} \end{bmatrix}$ 表示化窗的大小为 $3*3$, 权

重矩阵 $\begin{bmatrix} W_{11} & W_{12} & W_{13} \\ W_{21} & W_{22} & W_{23} \\ W_{31} & W_{32} & W_{33} \end{bmatrix}$ 的值为 $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$ 。

对于数据块 $\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$ 的最大池化，其计算结果应为：Max (a11,

a12, a13, a21, a22, a23, a31, a32, a33), Max () 表示求最大值。

如图 15 至图 21 所示，采用本申请的神经网络处理装置，执行输入特征
5 数据与权重的载入和最大池化计算的过程如下。

预先将权重 $\begin{bmatrix} W_{11} & W_{12} & W_{13} \\ W_{21} & W_{22} & W_{23} \\ W_{31} & W_{32} & W_{33} \end{bmatrix}$ (值为 $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$) 加载到脉动阵列中 (图 15

至图 21 中未示出)。

在第一个周期 (T1)，参见图 15 所示，输入特征数据 a11 进入计算单元 C00，其中，输入特征数据 a11 从计算单元 C00 的左边载入并从左向右流动。

10 在第一个周期结束时，计算单元 C00 的计算结果为 Max (a11)，即 a11。计算单元 C00 的计算结果 Max (a11) 从上向下流动。

在第二个周期 (T2)，参见图 16 所示。关于输入特征数据的载入与流动，以及计算结果的流动与图 9 类似，具体描述请参见上文结合图 9 的描述，为了简洁，这里不再赘述。在第二个周期结束时，计算单元 C00 的计算结果为
15 Max (a12)，计算单元 C01 的计算结果为 Max (a11)，计算单元 C10 的计算结果为 Max (a11, a21)。各个计算单元的计算结果从上向下流动。

在第三个周期 (T3)，参见图 17 所示。关于输入特征数据的载入与流动，以及计算结果的流动与图 10 类似，具体描述请参见上文结合图 10 的描述，为了简洁，这里不再赘述。在第三个周期结束时，计算单元 C00 的计算结果
20 为 Max (a13)，计算单元 C01 的计算结果为 Max (a12)，计算单元 C02 的计算结果为 Max (a11)，计算单元 C10 的计算结果为 Max (a12, a22)，计算单元 C11 的计算结果为 Max (a11, a21)，计算单元 C20 的计算结果为

Max (Max(a11, a21), a31)。各个计算单元的计算结果从上向下流动。

在第四个周期(T4), 参见图 18 所示。关于输入特征数据的载入与流动, 以及计算结果的流动与图 11 类似, 具体描述请参见上文结合图 11 的描述, 为了简洁, 这里不再赘述。在第四个周期结束时, 计算单元 C01 的计算结果
5 为 Max (a13), 计算单元 C02 的计算结果为 Max (a12), 计算单元 C10 的计算结果为 Max (a13, a23), 计算单元 C11 的计算结果为 Max (a12, a22), 计算单元 C12 的计算结果为 Max (a11, a21), 计算单元 C20 的计算结果为 Max (Max(a12, a22), a32), 计算单元 C21 的计算结果为 Max (Max(a11, a21), a31), 各个计算单元的计算结果从上向下流动。

10 在第五个周期(T5), 参见图 19 所示。关于输入特征数据的载入与流动, 以及计算结果的流动与图 12 类似, 具体描述请参见上文结合图 12 的描述, 为了简洁, 这里不再赘述。在第五个周期结束时, 计算单元 C02 的计算结果为 Max (a13), 计算单元 C11 的计算结果为 Max (a13, a23), 计算单元 C12 的计算结果为 Max (a12, a22), 计算单元 C20 的计算结果为 Max (Max(a13, a23), a33), 计算单元 C21 的计算结果为 Max (Max(a12, a22), a32), 计算
15 单元 C22 的计算结果为 Max (Max(a11, a21), a31), 各个计算单元的计算结果从上向下流动。

在第六个周期(T6), 参见图 20 所示。关于输入特征数据的载入与流动, 以及计算结果的流动与图 13 类似, 具体描述请参见上文结合图 13 的描述, 为了简洁, 这里不再赘述。在第六个周期结束时, 计算单元 C12 的计算结果
20 为 Max (a13, a23), 计算单元 C21 的计算结果为 Max(Max(a13,a23), a33), 计算单元 C22 的计算结果为 Max(Max(a12, a22), a32), 各个计算单元的计算结果从上向下流动。

在第七个周期(T7), 参见图 21 所示。关于输入特征数据的载入与流动, 以及计算结果的流动与图 14 类似, 具体描述请参见上文结合图 14 的描述, 为了简洁, 这里不再赘述。在第七个周期结束时, 计算单元 C22 的计算结果
25

为 $\text{Max}(\text{Max}(a_{13}, a_{23}), a_{33})$, 各个计算单元的计算结果从上向下流动。

可知, 第三个周期结束时计算单元 C20 的计算结果 $\text{Max}(\text{Max}(a_{11}, a_{21}), a_{31})$ 、第五个周期结束时计算单元 C21 的计算结果 $\text{Max}(\text{Max}(a_{12}, a_{22}), a_{32})$ 、以及第七个周期结束时计算单元 C22 的计算结果 $\text{Max}(\text{Max}(a_{13},$

5 $a_{23}), a_{33})$ 的最大值为 $\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$ 的最大池化操作的计算结果。

处理单元 130 用于接收计算单元 C20、C21 与 C22 输出的中间计算结果, 并对第三个周期结束时计算单元 C20 的计算结果、第五个周期结束时计算单元 C21 的计算结果、以及第七个周期结束时计算单元 C22 的计算结果作累

加, 得到数据块 $\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$ 的最大池化操作的计算结果。

10 上述可知, 本申请提供的基于脉动阵列的神经网络处理装置既可以实现卷积操作, 也可以实现池化操作 (平均池化、最大池化)。具体地, 卷积操作与池化操作可以共用一套数据通道, 通过控制信号的指示和/或修改权重值, 就可以使得同一脉动阵列和处理单元即可以实现卷积操作, 又可以实现池化操作。因此, 在本申请中, 池化操作与卷积操作复用一套硬件装置, 从而提
15 高硬件利用率, 节省成本。

应理解, 上述图 7 至图 21 仅为示例而非限定。实际应用中, 卷积核或者池化窗的大小不一定是 3×3 ; 输入特征矩阵 (即输入特征图) 的维度不一定是二维的, 还可以是三维的; 卷积核的维度也可能是 3 维的。此外, 实际应用中, 在一次卷积运算中, 可能针对多个 2 维的卷积核进行计算, 或者,
20 这都能对多个 3 维的卷积核进行计算。本申请提供的神经网络处理装置也可实现上述场景下的神经网络操作。

例如, 如图 22 所示, 采用本申请提供的神经网络处理装置实现针对 9 个 3×3 的卷积核的卷积操作。

权重映射方式为: 平铺映射。

25 如图 22 所示, 一个大小是 9×9 的脉动阵列可以刚好映射满 9 个 3×3 的滤波器 (filter) 的权重矩阵, 并且每个权重矩阵直接映射, 不需要翻转。

输入特征图中的输入特征数据载入脉动阵列的方式为：输入特征图中的每行输入特征数据按照顺序输入到脉动阵列的每行。每个周期，每行输入一个输入特征数据到脉动阵列的最左侧一列计算单元（MAC）中，同时每个计算单元把已经保存的输入特征数据向右传播到相邻的计算单元中。所以在列方向，每个计算单元接收到的输入特征数据来自于它左侧的计算单元。在行方向，相邻两行，下面一行总是延迟于上面一行一个周期。

计算时序可参考上文描述的图 7 至图 14 为了简洁，这里不再赘述。

如图 23 所示，本申请实施例还提供一种基于脉动阵列的池化处理装置 2300。该装置 2300 用于执行平均池化或最大池化。该装置 2300 包括控制单元 2310、脉动阵列 2320、处理单元 2330。

控制单元 2310，用于向脉动阵列 2320 和处理单元 2330 发送控制信号，控制信号指示平均池化操作或最大化操作。

脉动阵列 2320，包括多个计算单元 3000。脉动阵列 2320 中的每个计算单元 3000 的内部结构相同。

脉动阵列 2320 中的各个计算单元 3000 用于，接收控制信号、输入特征数据以及来自沿脉动阵列 2320 的第二维度的上一级计算单元的第一中间处理结果；根据控制信号，对输入特征数据以及第一中间处理结果进行池化相关操作，获得第二中间处理结果；向沿着脉动阵列 2320 的第二维度的下一级计算单元发送第二中间处理结果；向沿着脉动阵列 2320 的第一维度的下一级计算单元发送输入特征数据。

如图 23 所示，对于脉动阵列 2320 中沿着第一维度上的输入端的计算单元来说，其可以从脉动阵列 2320 之外的特征数据存储器 2340 中接收输入特征值。其余计算单元均是从沿着第一维度的上一级计算单元接收输入特征值。

可选地，由主处理器确定需要载入脉动阵列 2320 中的输入特征数据，然后将该输入特征数据从特征数据存储器 2340 载入到脉动阵列 2320 中。其中，该主处理器可以为控制单元 2310，也可以为神经网络系统中的其他处理器。

可选地，本申请提供的神经网络处理装置 2300 包括特征数据存储器 2340。

可选地，本申请提供的神经网络处理装置 2300 不包括特征数据存储器 2340，但是包括用于与特征数据存储器 2340。

处理单元 2330 与脉动阵列 2320 输出端的计算单元 3000 连接，处理单元 2230 用于接收控制信号以及脉动阵列输出端的计算单元输出的第二中间处理结果，根据控制信号对第二中间处理结果进行池化相关操作，获得池化输出结果。

- 5 脉动阵列 2320 输出端的计算单元指的是，脉动阵列 2320 中最终输出中间处理结果的计算单元，例如，在图 2300 中，指的是，脉动阵列 2320 的最后一行的计算单元。

例如，脉动阵列 2320 输出端的计算单元的发送接口与处理单元 2330 的接收接口通信连接。

- 10 当控制信号指示平均池化操作时，处理单元 2330 对第二中间处理结果进行求平均运算。当控制信号指示最大池化操作时，处理单元 2330 对第二中间处理结果进行比较运算。

上述可知，控制单元 2310 用于控制脉动阵列 2320 和处理单元 2330 进行平均池化操作或最大池化操作，可以在一定程度上提高硬件利用率。

- 15 此外，利用脉动阵列实现池化运算，可以有效降低池化操作过程中读数据的次数。

因此，本申请提供的池化处理装置，可以通过使得平均池化操作和最大池化操作复用一套硬件装置，从而可以提高硬件利用率，节省成本，同时也可以有效降低池化操作过程中读数据的次数。

- 20 计算单元 3000 的内部结构可以有多种实现方式。

可选地，作为一种实现方式，各个计算单元 3000 的内部结构如下。

特征寄存器，用于存储输入特征数据，还用于向沿着脉动阵列 2320 的第一维度的后一级计算单元转发输入特征数据。

加法器，与特征寄存器连接。

- 25 比较器，与特征寄存器连接。

在控制信号指示平均池化操作时，加法器用于，接收输入特征数据与来自沿着第二维度的前一级计算单元的第一中间处理结果，对输入特征数据与第一中间处理结果进行累加，获得第二中间处理结果，向沿着第二维度的后一级计算单元发送第二中间处理结果。

- 30 在控制信号指示最大池化操作时，比较器用于接收输入特征数据与第一中间结果，并输出输入特征数据与第一中间结果中的最大值，向沿着第二维

度的后一级计算单元发送第二中间处理结果，第二中间处理结果为最大值。

可选地，作为另一种实现方式，各个计算单元 3000 的内部结构如下。

特征寄存器，用于存储输入特征数据，还用于向沿着脉动阵列 2320 的第一维度的后一级计算单元转发输入特征数据。

5 加法器，与特征寄存器连接。

选择器，与特征寄存器和加法器连接。

在控制信号指示平均池化操作时，加法器用于，接收输入特征数据与来自沿着第二维度的前一级计算单元的第一中间处理结果，对输入特征数据与第一中间处理结果进行累加，获得第二中间处理结果，向沿着第二维度的后

10 一级计算单元发送第二中间处理结果。

在控制信号指示最大池化操作时，加法器用于，接收输入特征数据与第一中间处理结果，输出乘积与第一中间处理结果的相减结果；选择器用于接收输入特征数据、相减结果以及第一中间处理结果，根据相减结果从输入特征数据与第一中间处理结果中选择出最大值，并向沿着第二维度的后一级计

15 算单元发送第二中间处理结果，第二中间处理结果为最大值。

可选地，在本实施例中，处理单元 2330 用于：在控制信号指示平均池化操作时，基于池化窗的大小，对接收的第二中间处理结果进行求平均，获得平均池化操作的输出结果；或在控制信号指示最大池化操作时，对接收的第二中间处理结果进行比较，获得最大池化操作的输出结果。

20 因此，本申请提供的神经网络处理装置 2300，可以通过使得平均池化操作和最大池化操作复用一套硬件装置，从而可以提高硬件利用率，节省成本，同时也可以有效降低池化操作过程中读数据的次数。

需要说明的是，虽然上文按照特定顺序描述了各个步骤，但是并不意味着必须按照上述特定顺序来执行各个步骤，实际上，这些步骤中的一些可以

25 并发执行，甚至改变顺序，只要能够实现所需要的功能即可，本申请对各个步骤的执行顺序不做严格限定。

本申请实施例提供的神经网络处理装置可以在现场可编程门阵列（field-programmable gate array, FPGA）或专门应用的集成电路（application specific integrated circuit, ASIC）上实现。

30 本发明实施例还提供了一种计算单元的控制方法。该计算单元例如可以是上文任一实施例描述的计算机单元 200。该控制方法由上述实施例描述的控

制单元 110 执行。该控制方法可以包括如上文中描述的由控制单元 110 执行的
5 的处理流程，为避免重复，此处不再详述。

在上述实施例中，可以全部或部分地通过软件、硬件、固件或者其他任
意组合来实现。当使用软件实现时，可以全部或部分地以计算机程序产品的
5 形式实现。所述计算机程序产品包括一个或多个计算机指令。在计算机上加
载和执行所述计算机程序指令时，全部或部分地产生按照本发明实施例所述
的流程或功能。所述计算机可以是通用计算机、专用计算机、计算机网络、
或者其他可编程装置。所述计算机指令可以存储在计算机可读存储介质中，
或者从一个计算机可读存储介质向另一个计算机可读存储介质传输，例如，
10 所述计算机指令可以从一个网站站点、计算机、服务器或数据中心通过有线
(例如同轴电缆、光纤、数字用户线 (digital subscriber line, DSL)) 或无线
(例如红外、无线、微波等) 方式向另一个网站站点、计算机、服务器或数
据中心进行传输。所述计算机可读存储介质可以是计算机能够存取的任何可
用介质或者是包含一个或多个可用介质集成的服务器、数据中心等数据存储
15 设备。所述可用介质可以是磁性介质 (例如，软盘、硬盘、磁带)、光介质
(例如数字视频光盘 (digital video disc, DVD))、或者半导体介质 (例如固
态硬盘 (solid state disk, SSD)) 等。

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、或者计算机软件和电子硬件的结
20 合来实现。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。

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
25 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。另一点，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，可以是电性，机械或其它的形式。

30 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实施例方案的目的。

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计算单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。

5

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，可轻易想到变化或替换，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所述权利要求的保护范围为准。

10

权利要求

1. 一种神经网络处理装置，其特征在于，包括：

脉动阵列，包括多个计算单元；

控制单元，用于发送控制信号，所述控制信号指示如下神经网络操作中

5 的任一种：卷积操作、平均池化操作以及最大化操作；

所述脉动阵列中的各个计算单元用于，接收所述控制信号、输入特征数据、权重以及来自沿所述脉动阵列的第二维度的上一级计算单元的第一中间处理结果；根据所述控制信号，对所述输入特征数据、所述权重以及所述第一中间处理结果进行神经网络相关操作，获得第二中间处理结果；向沿着所述脉动阵列的所述第二维度的下一级计算单元发送所述第二中间处理结果；向沿着所述脉动阵列的第一维度的下一级计算单元发送所述输入特征数据；

10 处理单元，所述处理单元与所述脉动阵列输出端的计算单元连接，所述处理单元用于接收所述控制信号以及所述脉动阵列输出端的计算单元输出的第二中间处理结果，根据所述控制信号对所述第二中间处理结果进行神经网络相关操作，获得神经网络输出结果。

2. 根据权利要求1所述的装置，其特征在于，在所述控制信号指示最大池化操作或平均池化操作时，所述权重的值为1。

3. 根据权利要求1所述的装置，其特征在于，在所述控制信号指示卷积操作时，所述权重在所述脉动阵列中的分布方式与所述权重在卷积核中的分布方式一致或呈镜像。

4. 根据权利要求1至3中任一项所述的装置，其特征在于，所述装置包括：

特征数据存储器，用于存储输入特征数据；

权重存储器，用于存储权重；

25 所述控制单元用于，将所述特征数据存储器中的输入特征数据载入所述脉动阵列，将所述权重存储器中的权重载入所述脉动阵列。

5. 根据权利要求4所述的装置，其特征在于，所述装置还包括：

另一权重存储器，用于在所述控制信号指示卷积操作时，预存下一次卷积操作的权重。

30 6. 根据权利要求1至5中任一项所述的装置，其特征在于，所述各个计算单元包括：

权重寄存器，用于存储权重；

特征寄存器，用于存储输入特征数据，还用于向沿着所述脉动阵列的所述第一维度的后一级计算单元转发所述输入特征数据；

乘法器，所述乘法器与所述权重寄存器和所述特征寄存器连接，所述乘法器用于输出所述权重与所述输入特征数据的乘积；

加法器，与所述乘法器里连接；

在所述控制信号指示卷积操作或平均池化操作时，所述加法器用于，接收所述乘积与来自沿着所述第二维度的前一级计算单元的第一中间处理结果，对所述乘积与所述第一中间处理结果进行累加，获得第二中间处理结果，向沿着所述第二维度的后一级计算单元发送所述第二中间处理结果。

7. 根据权利要求6所述的装置，其特征在于，

在所述控制信号指示最大池化操作时，所述加法器用于，接收所述乘积与所述第一中间处理结果，输出所述乘积与所述第一中间处理结果的相减结果；

15 所述计算单元还包括：

选择器，与所述加法器和所述乘法器连接，在所述控制信号指示最大池化操作时，所述选择器用于接收所述乘积、所述相减结果、以及所述第一中间处理结果，根据所述相减结果从所述乘积与所述第一中间处理结果中选择出最大值，并向沿着所述第二维度的后一级计算单元发送第二中间处理结果，所述第二中间处理结果为所述最大值。

8. 根据权利要求6所述的装置，其特征在于，所述计算单元还包括：

比较器，与所述乘法器连接；

在所述控制信号指示最大池化操作时，所述比较器用于接收所述乘积与所述第一中间结果，并输出所述乘积与所述第一中间结果中的最大值，向沿着所述第二维度的后一级计算单元发送第二中间处理结果，所述第二中间处理结果为所述最大值。

9. 根据权利要求1至8中任一项所述的装置，其特征在于，所述处理单元用于：

在所述控制信号指示卷积操作时，对接收的第二中间处理结果进行累加，获得所述卷积操作的输出结果；或

在所述控制信号指示平均池化操作时，基于池化窗的大小，对接收的第

二中间处理结果进行求平均，获得所述平均池化操作的输出结果；或

在所述控制信号指示最大池化操作时，对接收的第二中间处理结果进行比较，获得所述最大池化操作的输出结果。

10、根据权利要求 1 至 9 中任一项所述的装置，其特征在于，所述第一
5 维度对应于所述脉动阵列的行，所述第二维度对应于所述脉动阵列的列。

11. 一种池化处理装置，其特征在于，包括：

脉动阵列，包括多个计算单元；

控制单元，用于发送控制信号，所述控制信号指示平均池化操作或最大
化操作；

10 所述脉动阵列中的各个计算单元用于，接收所述控制信号、输入特征数
据以及来自沿所述脉动阵列的第二维度的上一级计算单元的第一中间处理
结果；根据所述控制信号，对所述输入特征数据以及所述第一中间处理结果
进行池化相关操作，获得第二中间处理结果；向沿着所述脉动阵列的所述第
二维度的下一级计算单元发送所述第二中间处理结果；向沿着所述脉动阵列
15 的第一维度的下一级计算单元发送所述输入特征数据；

处理单元，与所述脉动阵列输出端的计算单元连接，所述处理单元用于
接收所述控制信号以及所述脉动阵列输出端的计算单元输出的第二中间处
理结果，根据所述控制信号对所述第二中间处理结果进行池化相关操作，获
得池化输出结果。

20 12.根据权利要求 11 所述的装置，其特征在于，所述装置还包括：

特征数据存储器，用于存储输入特征数据；

所述控制单元用于，将所述特征数据存储器中的输入特征数据载入所述
脉动阵列。

25 13. 根据权利要求 11 或 12 所述的装置，其特征在于，所述各个计算单
元包括：

特征寄存器，用于存储输入特征数据，还用于向沿着所述脉动阵列的所
述第一维度的后一级计算单元转发所述输入特征数据；

加法器，与所述特征寄存器连接；

30 在所述控制信号指示平均池化操作时，所述加法器用于，接收所述输入
特征数据与来自沿着所述第二维度的前一级计算单元的第一中间处理结果，
对所述输入特征数据与所述第一中间处理结果进行累加，获得第二中间处理

结果，向沿着所述第二维度的后一级计算单元发送所述第二中间处理结果。

14. 根据权利要求 13 所述的装置，其特征在于，

在所述控制信号指示最大池化操作时，所述加法器用于，接收所述输入特征数据与所述第一中间处理结果，输出所述乘积与所述第一中间处理结果的相减结果；

所述计算单元还包括：

选择器，与所述特征寄存器和所述加法器连接；

在所述控制信号指示最大池化操作时，所述选择器用于接收所述输入特征数据、所述相减结果以及所述第一中间处理结果，根据所述相减结果从所述输入特征数据与所述第一中间处理结果中选择出最大值，并向沿着所述第二维度的后一级计算单元发送第二中间处理结果，所述第二中间处理结果为所述最大值。

15. 根据权利要求 13 所述的装置，其特征在于，所述计算单元还包括：比较器，与所述特征寄存器连接；

在所述控制信号指示最大池化操作时，所述比较器用于接收所述输入特征数据与所述第一中间结果，并输出所述输入特征数据与所述第一中间结果中的最大值，向沿着所述第二维度的后一级计算单元发送第二中间处理结果，所述第二中间处理结果为所述最大值。

16. 根据权利要求 11 至 15 中任一项所述的装置，其特征在于，所述处理单元用于：

在所述控制信号指示平均池化操作时，基于池化窗的大小，对接收的第二中间处理结果进行求平均，获得所述平均池化操作的输出结果；或

在所述控制信号指示最大池化操作时，对接收的第二中间处理结果进行比较，获得所述最大池化操作的输出结果。

17. 根据权利要求 11 至 16 中任一项所述的装置，其特征在于，所述第一维度对应于所述脉动阵列的行，所述第二维度对应于所述脉动阵列的列。

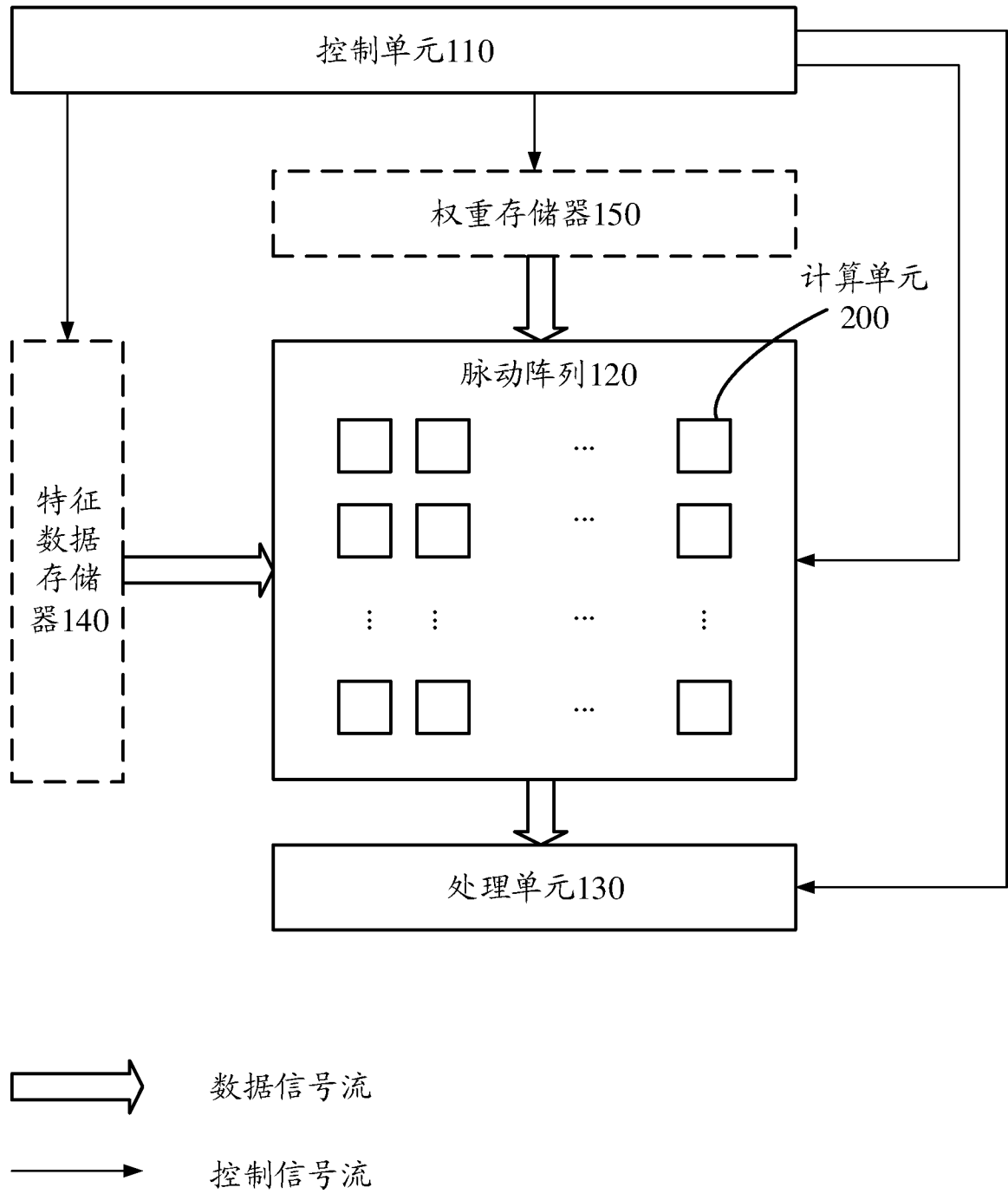


图 1

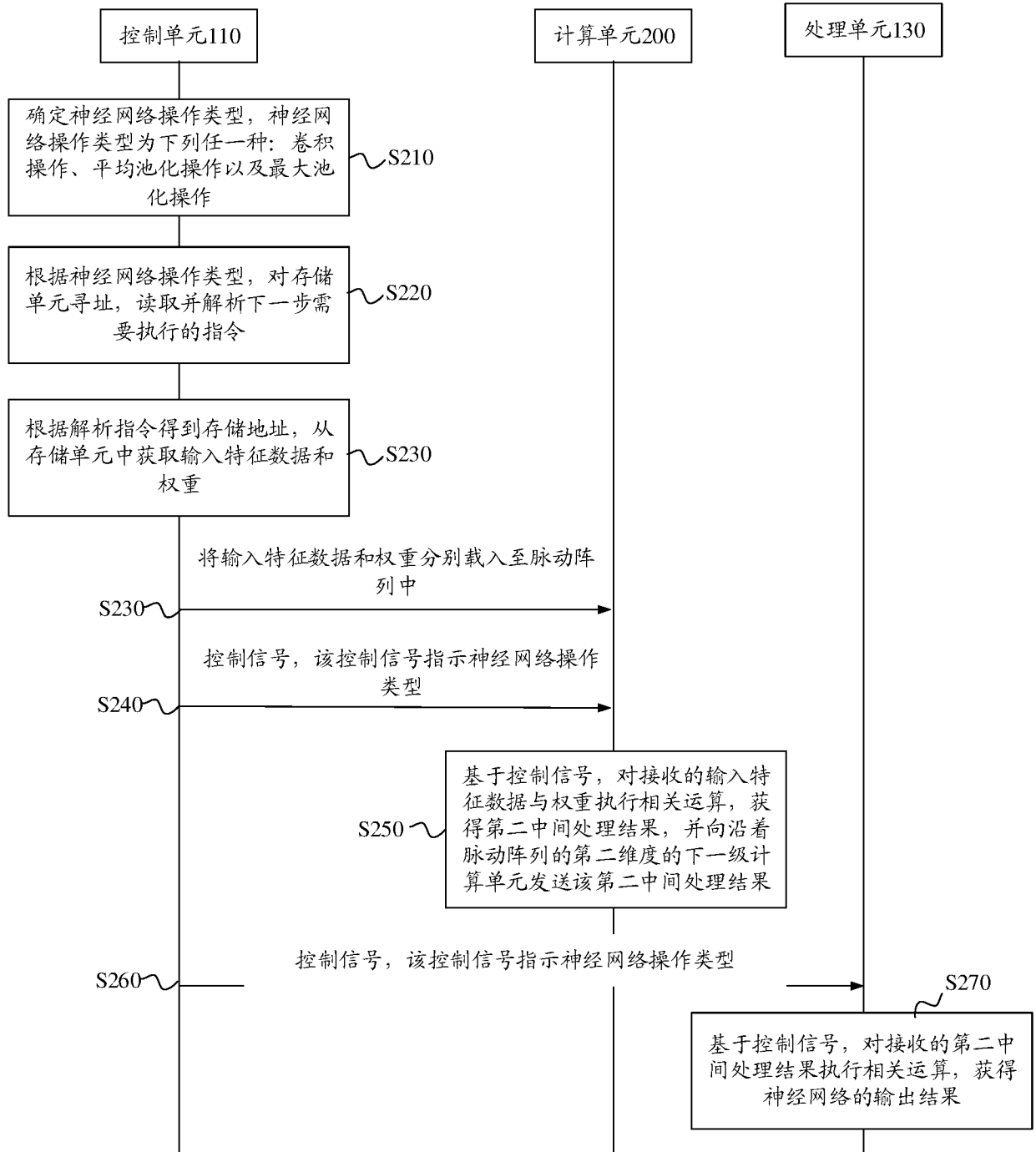


图 2

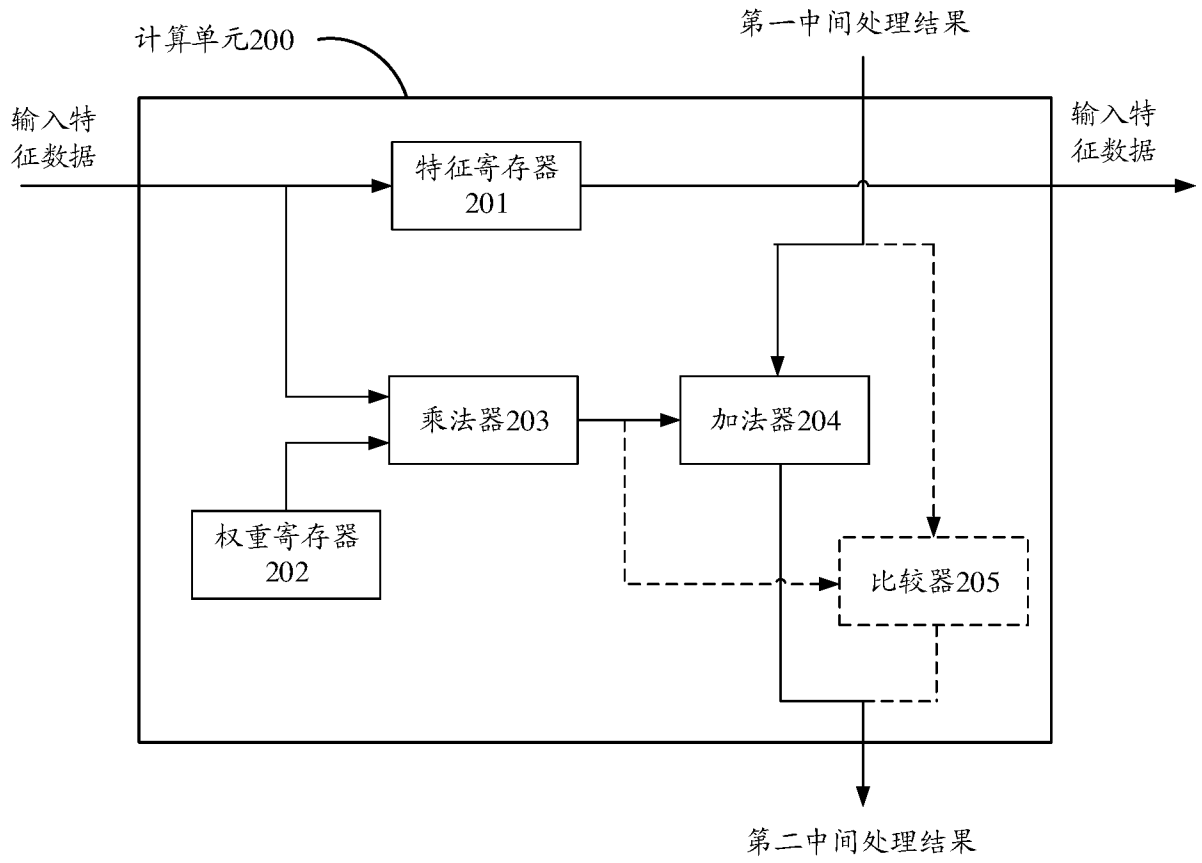


图 3

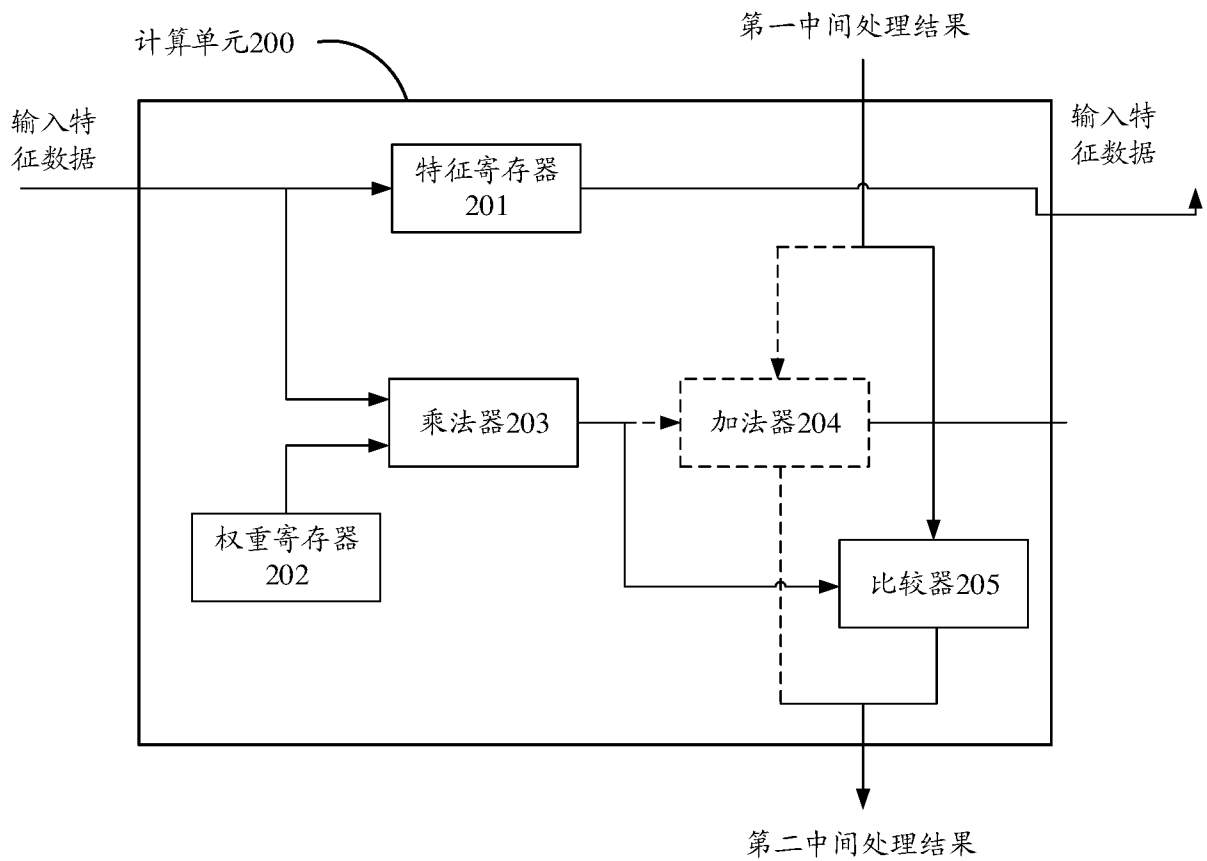


图 4

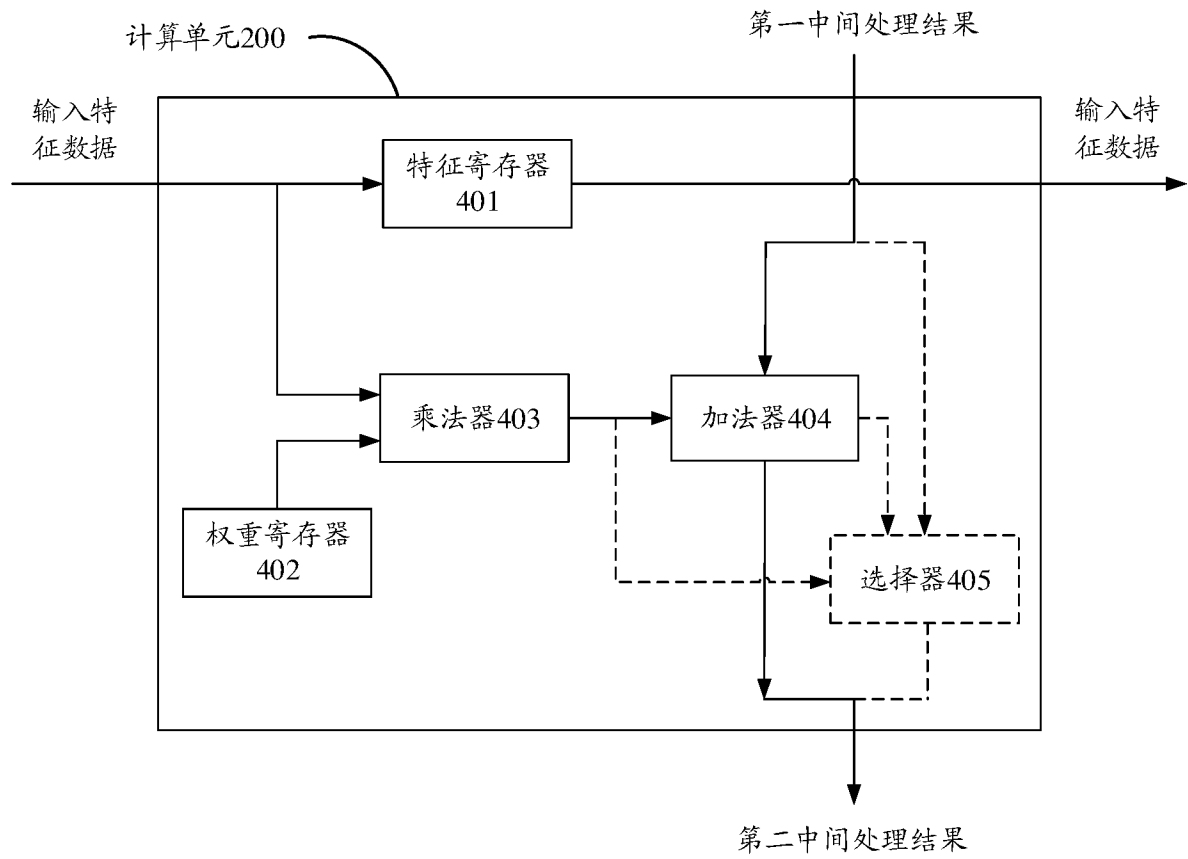


图 5

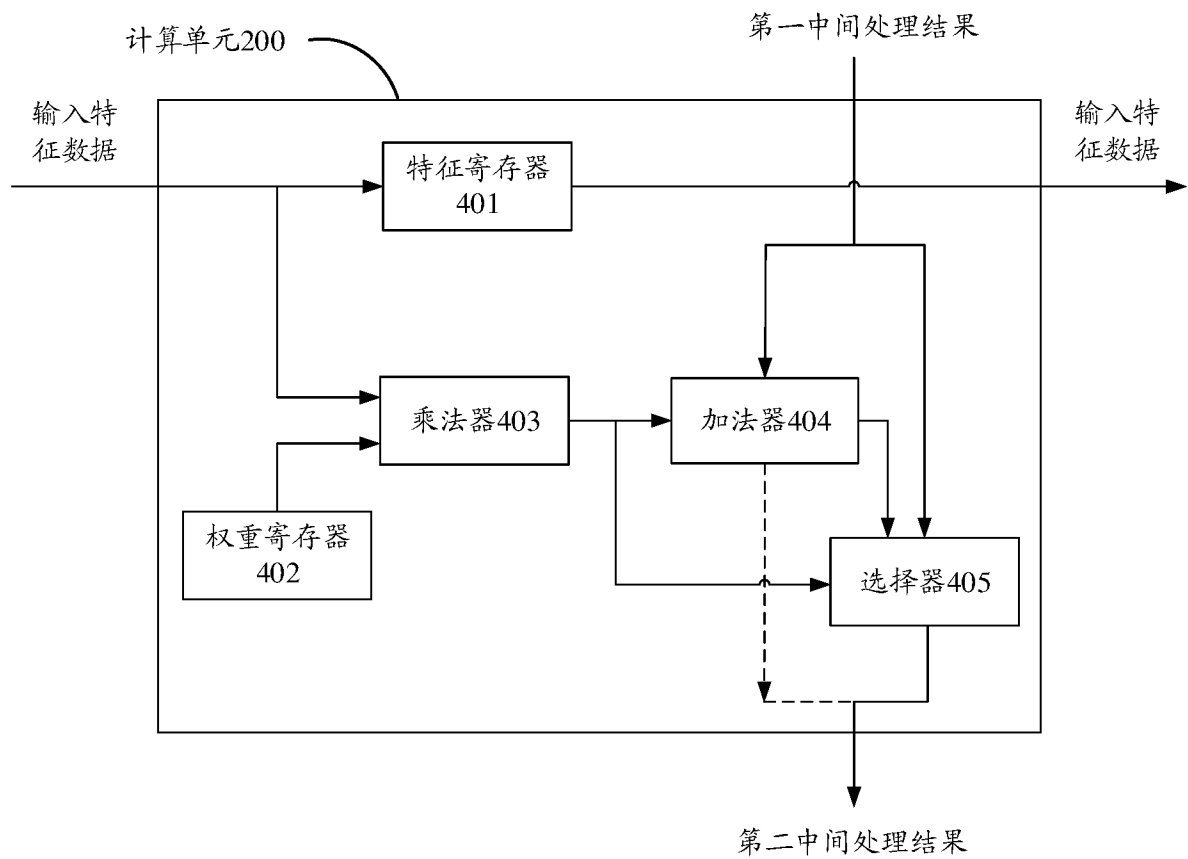


图 6

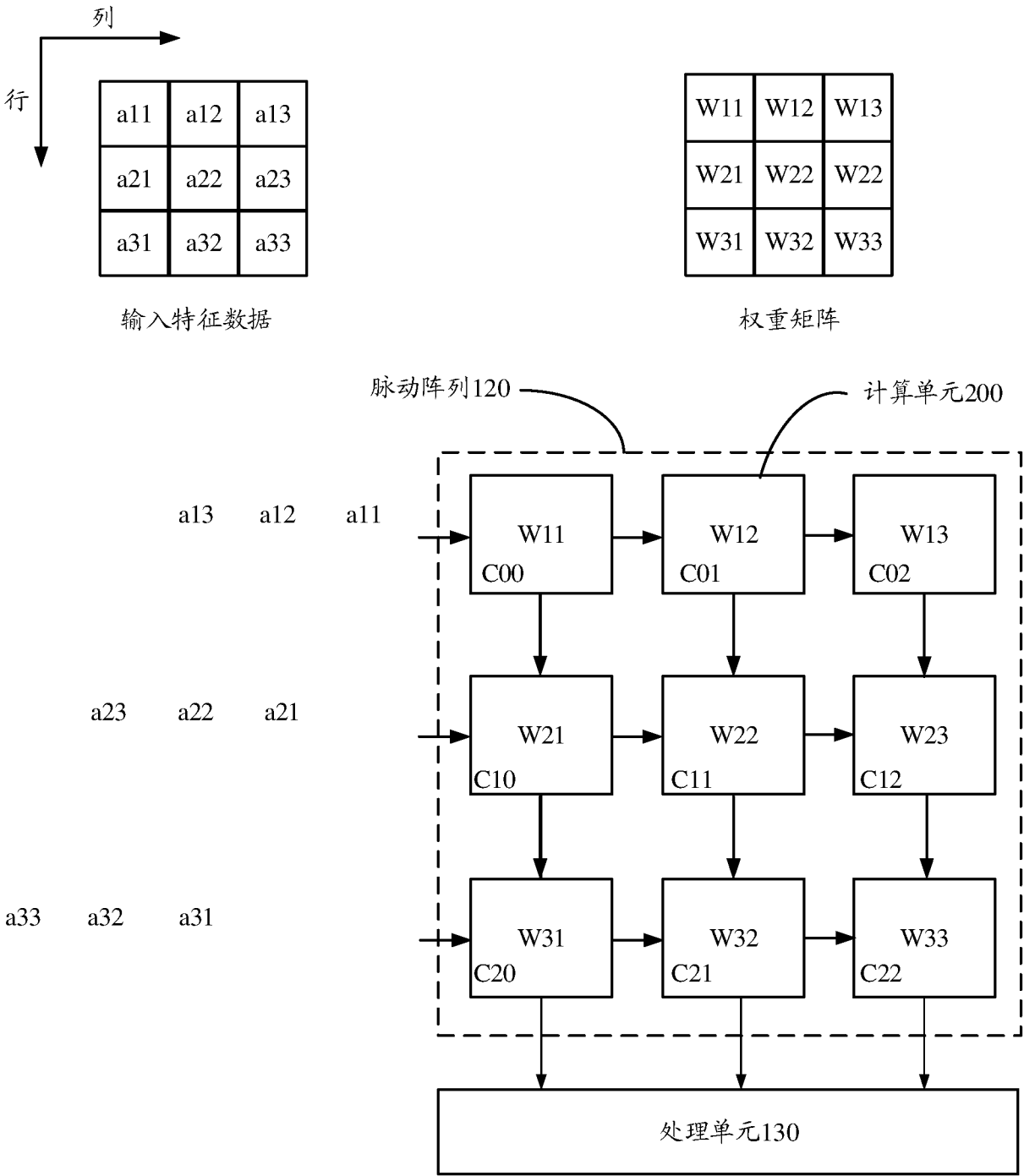
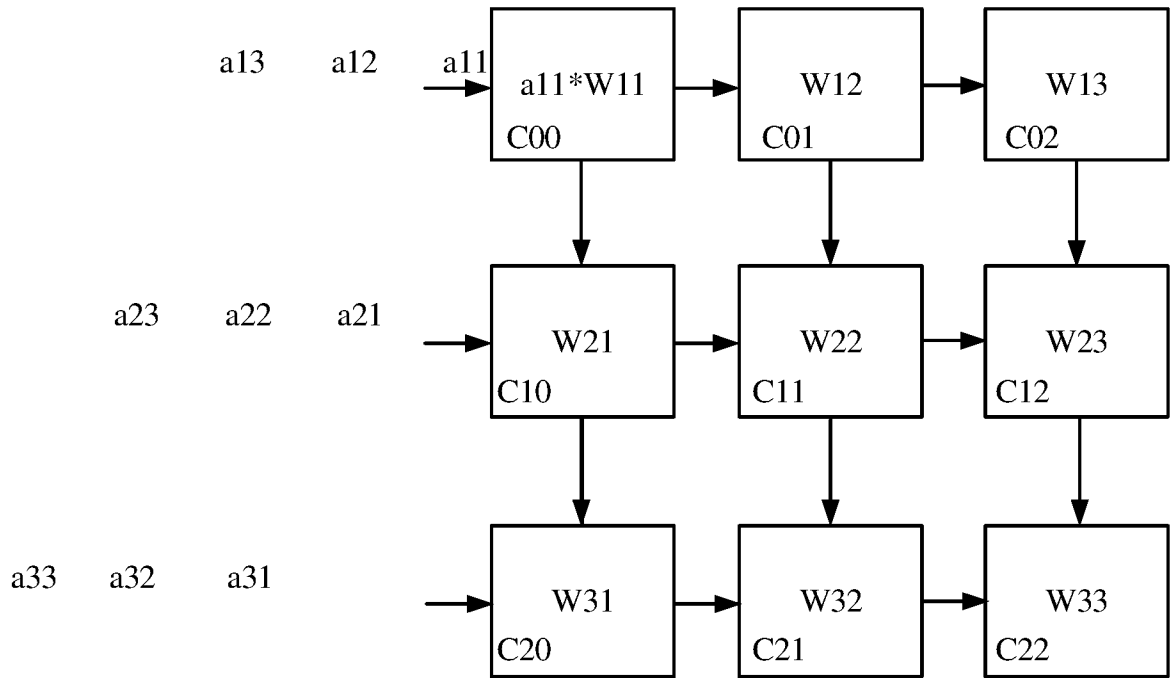
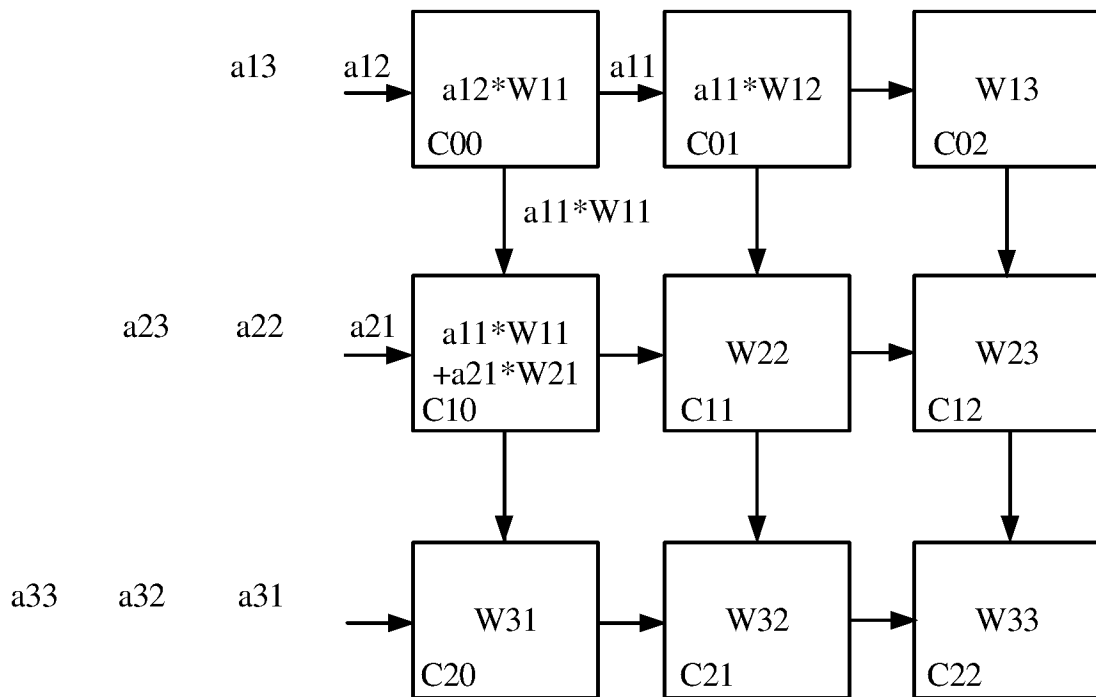


图 7



T1

图 8



T2

图 9

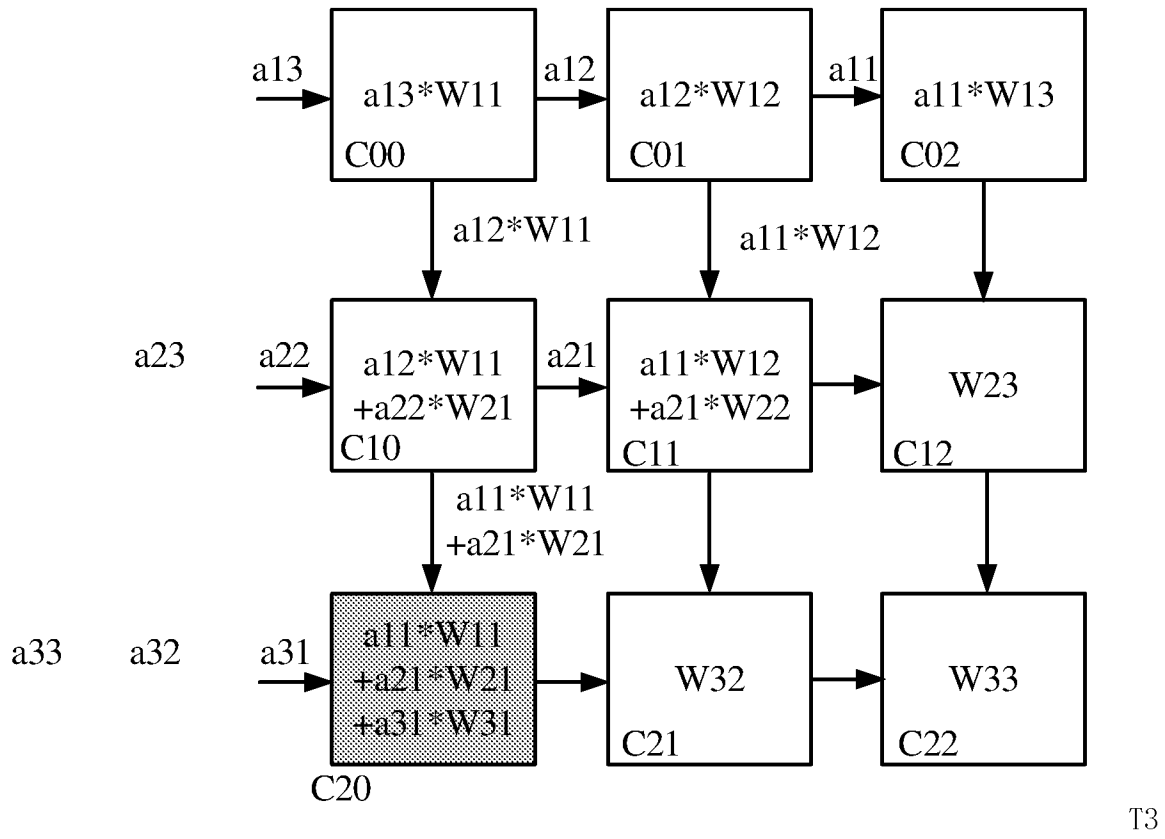


图 10

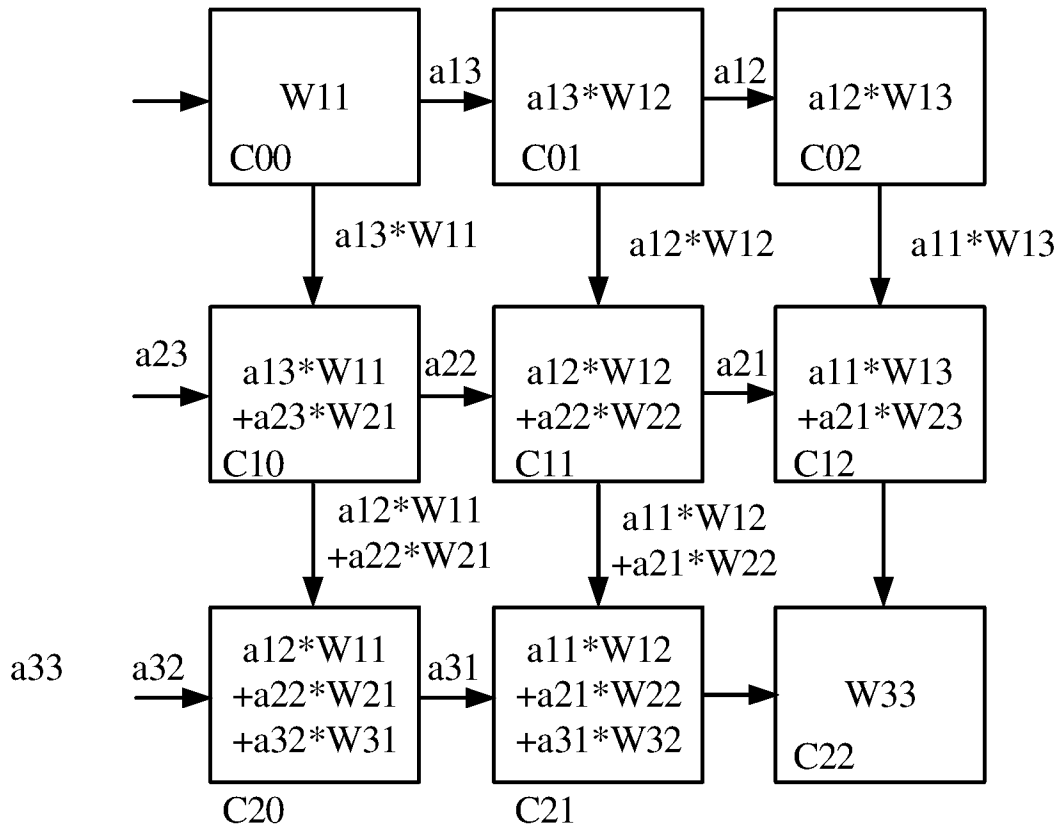
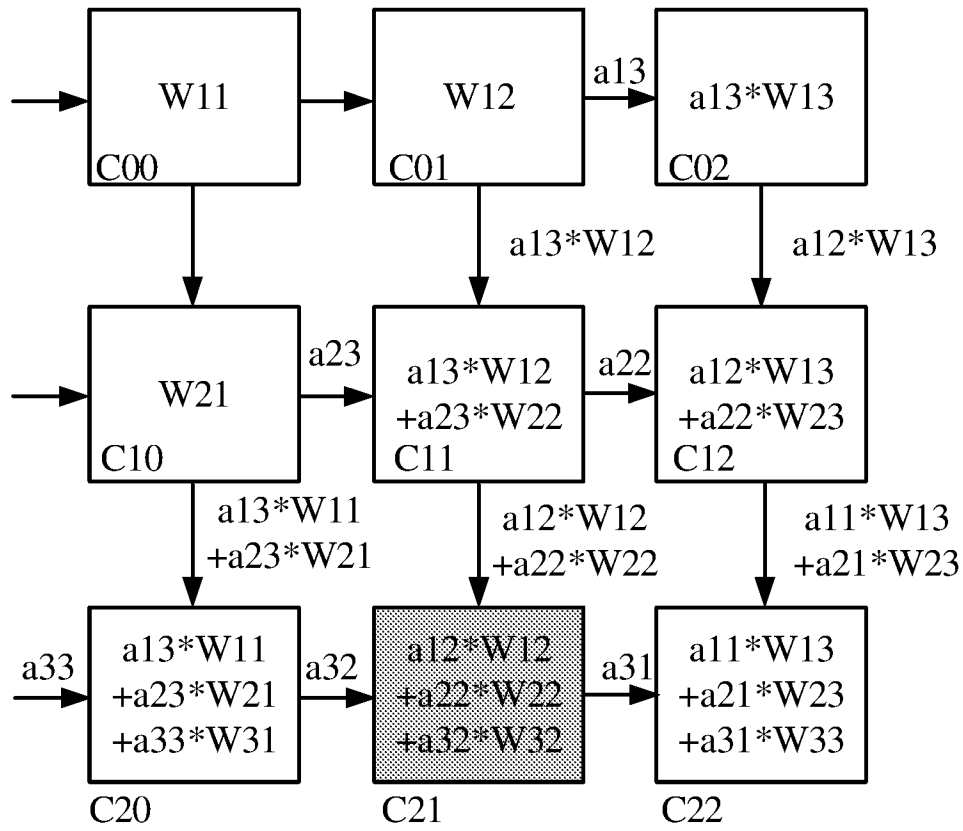
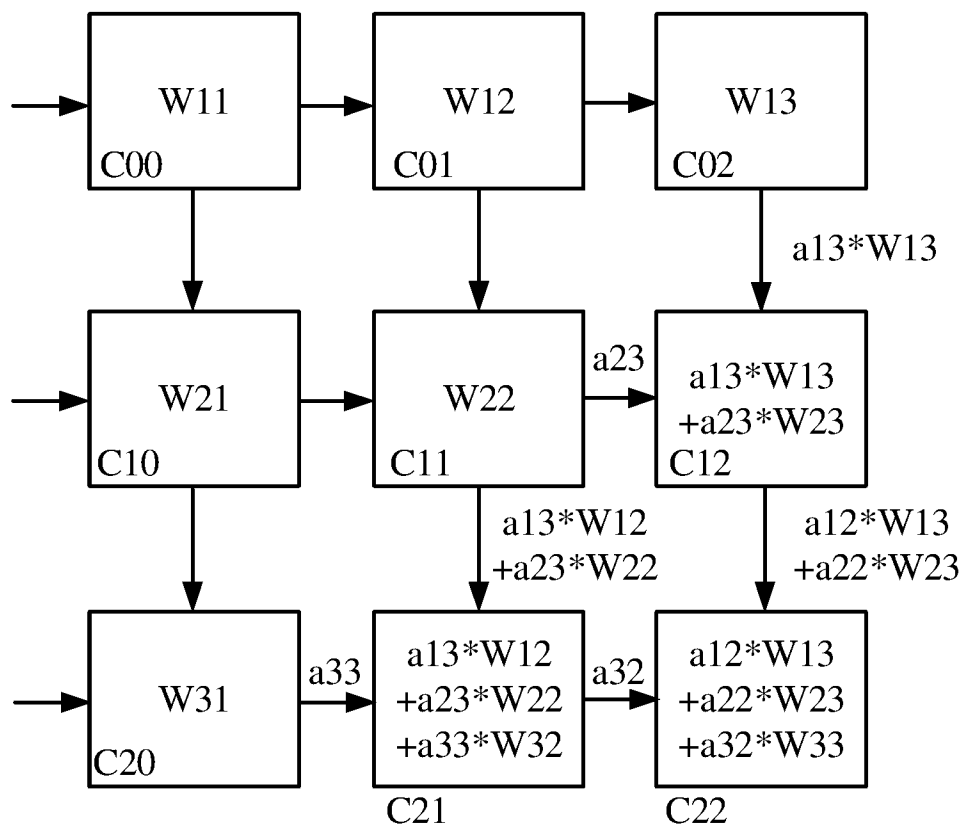


图 11



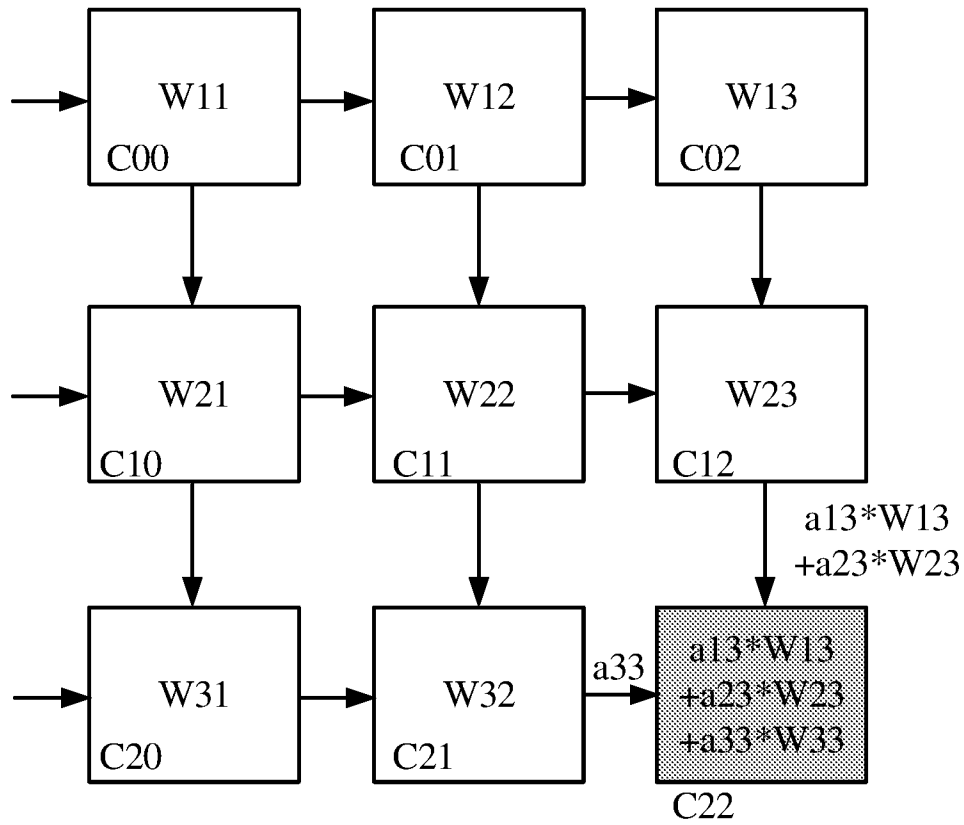
T5

图 12



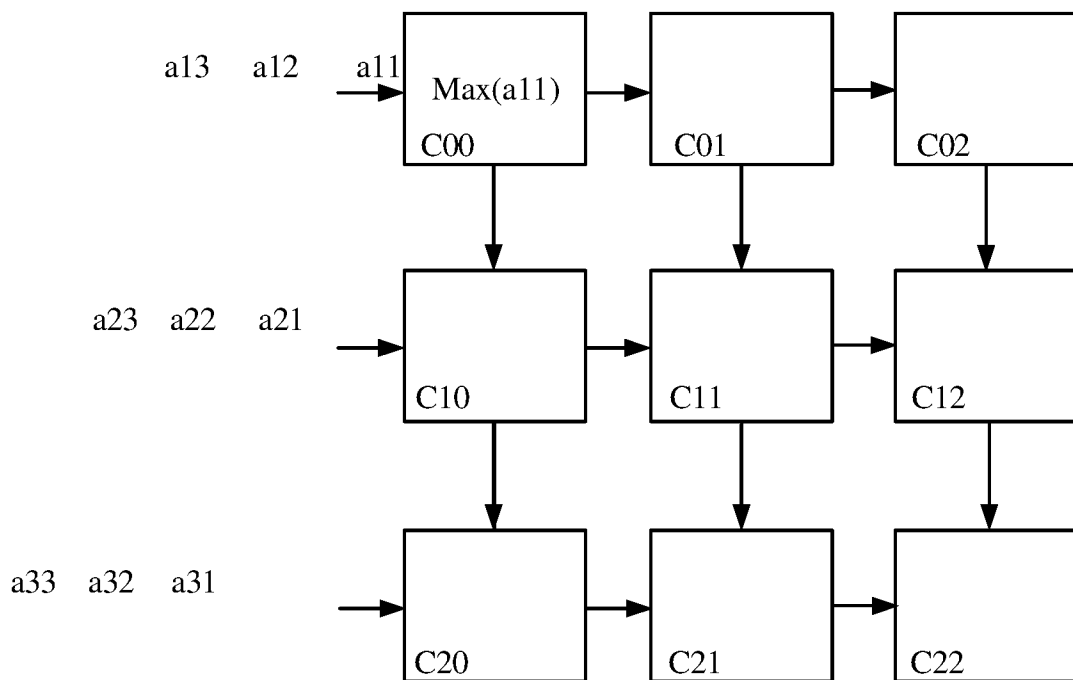
T6

图 13



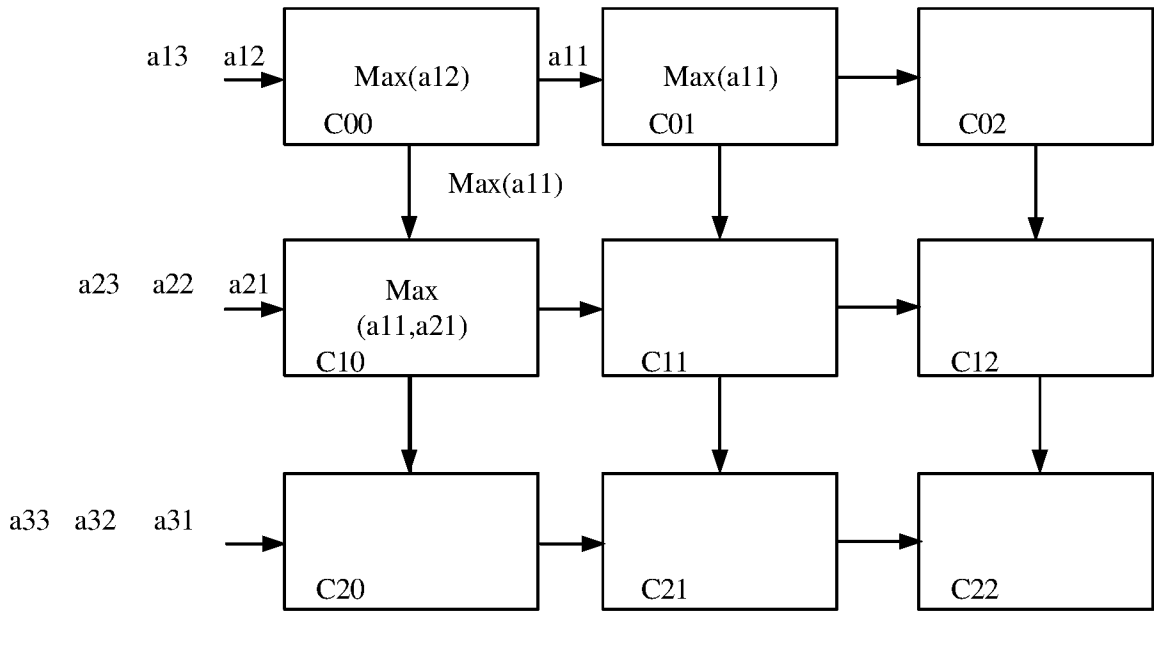
T7

图 14



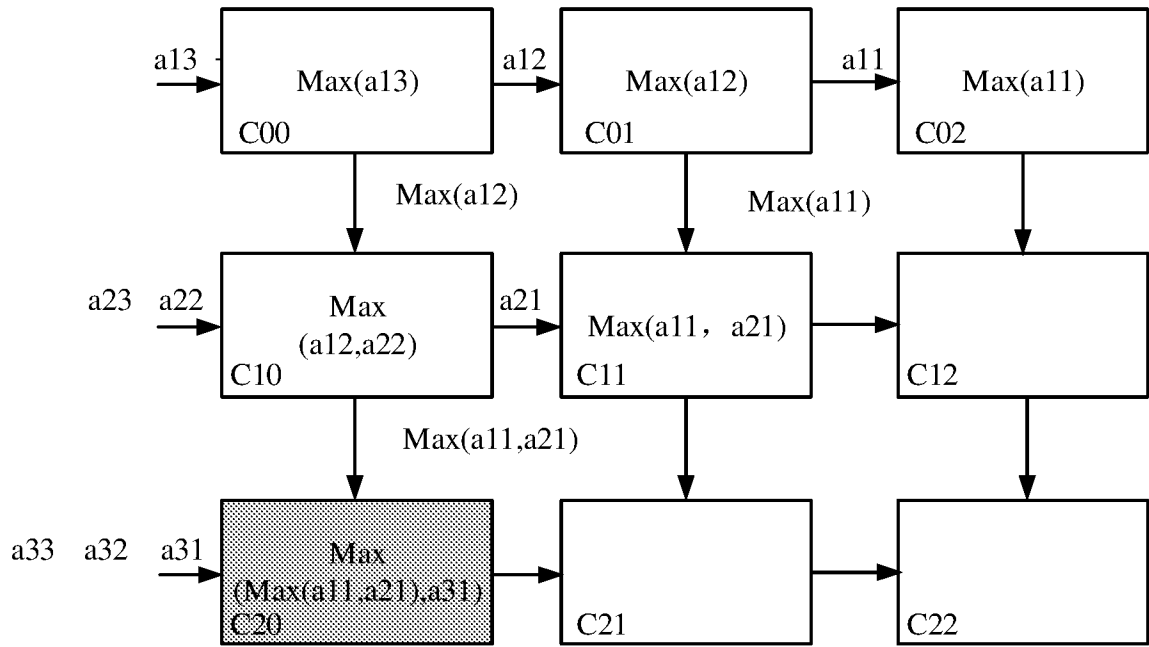
T1

图 15



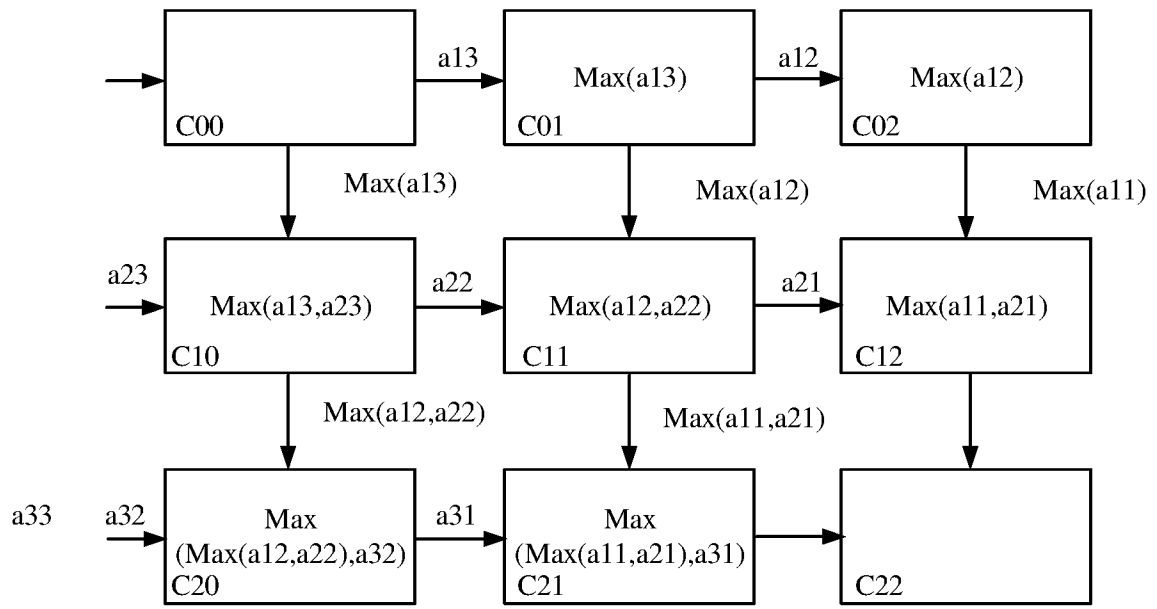
T2

图 16



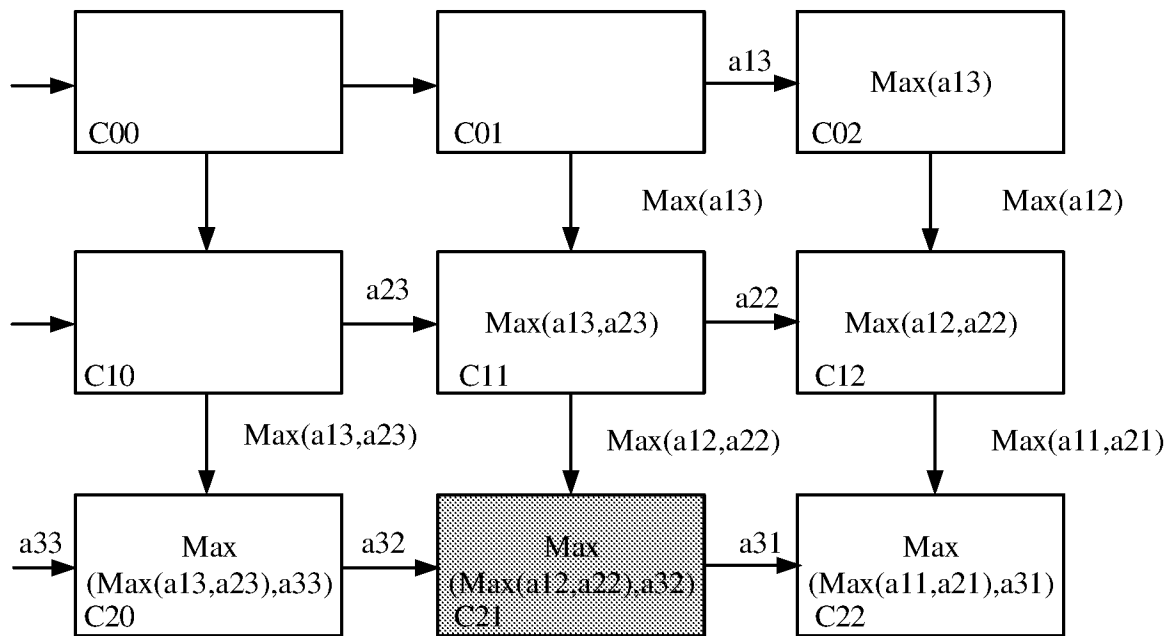
T3

图 17



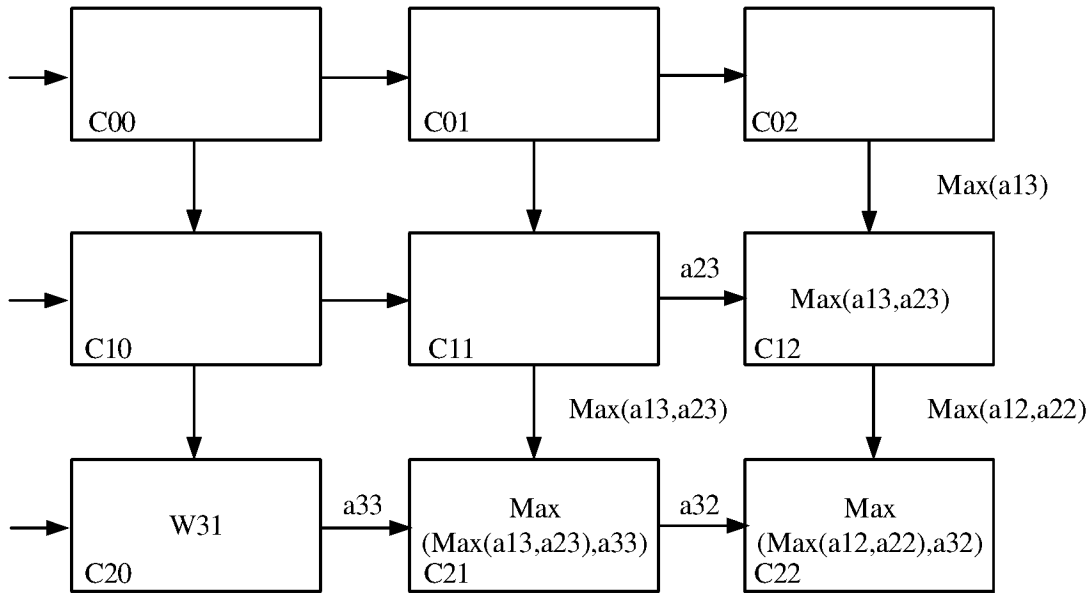
T4

图 18



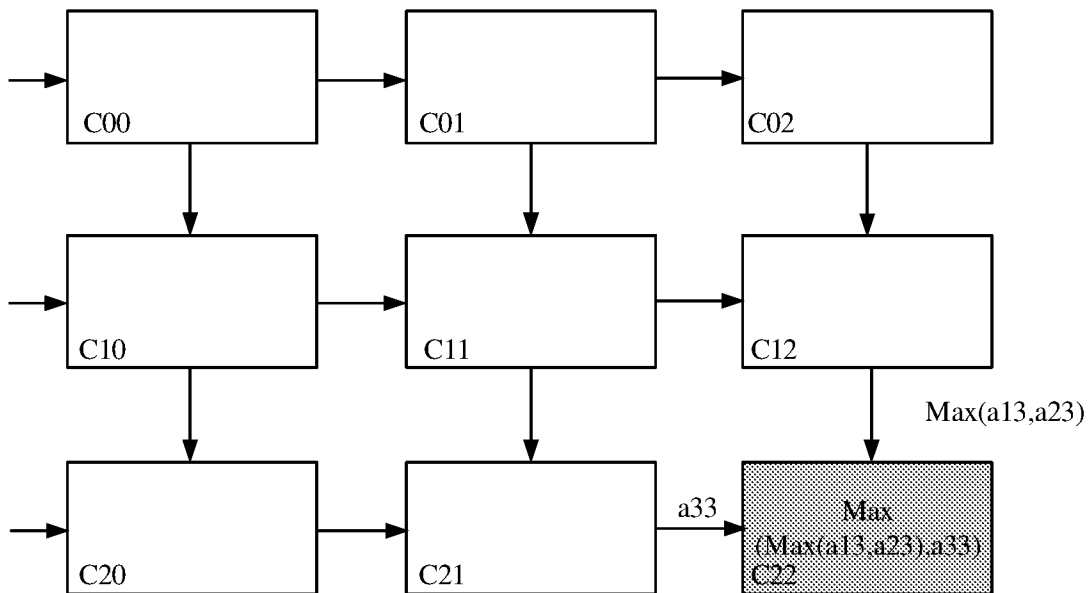
T5

图 19



T6

图 20



T7

图 21

脉动阵列

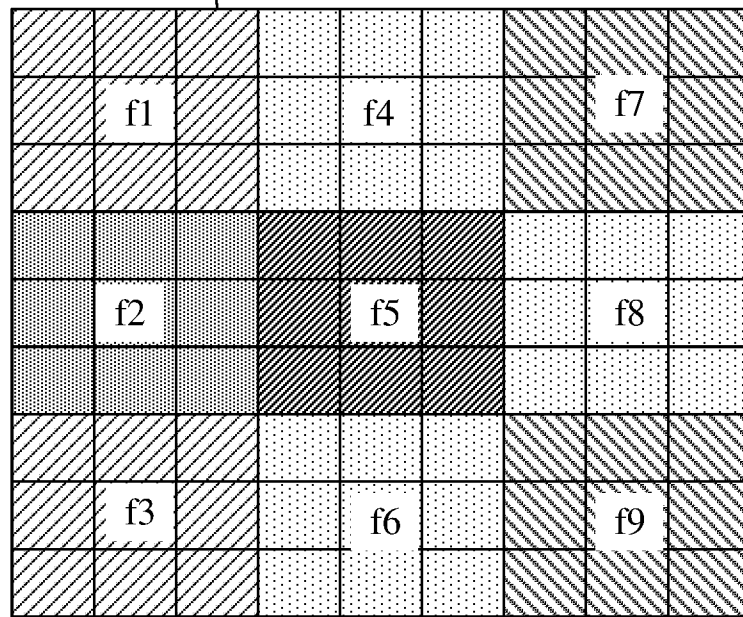


图 22

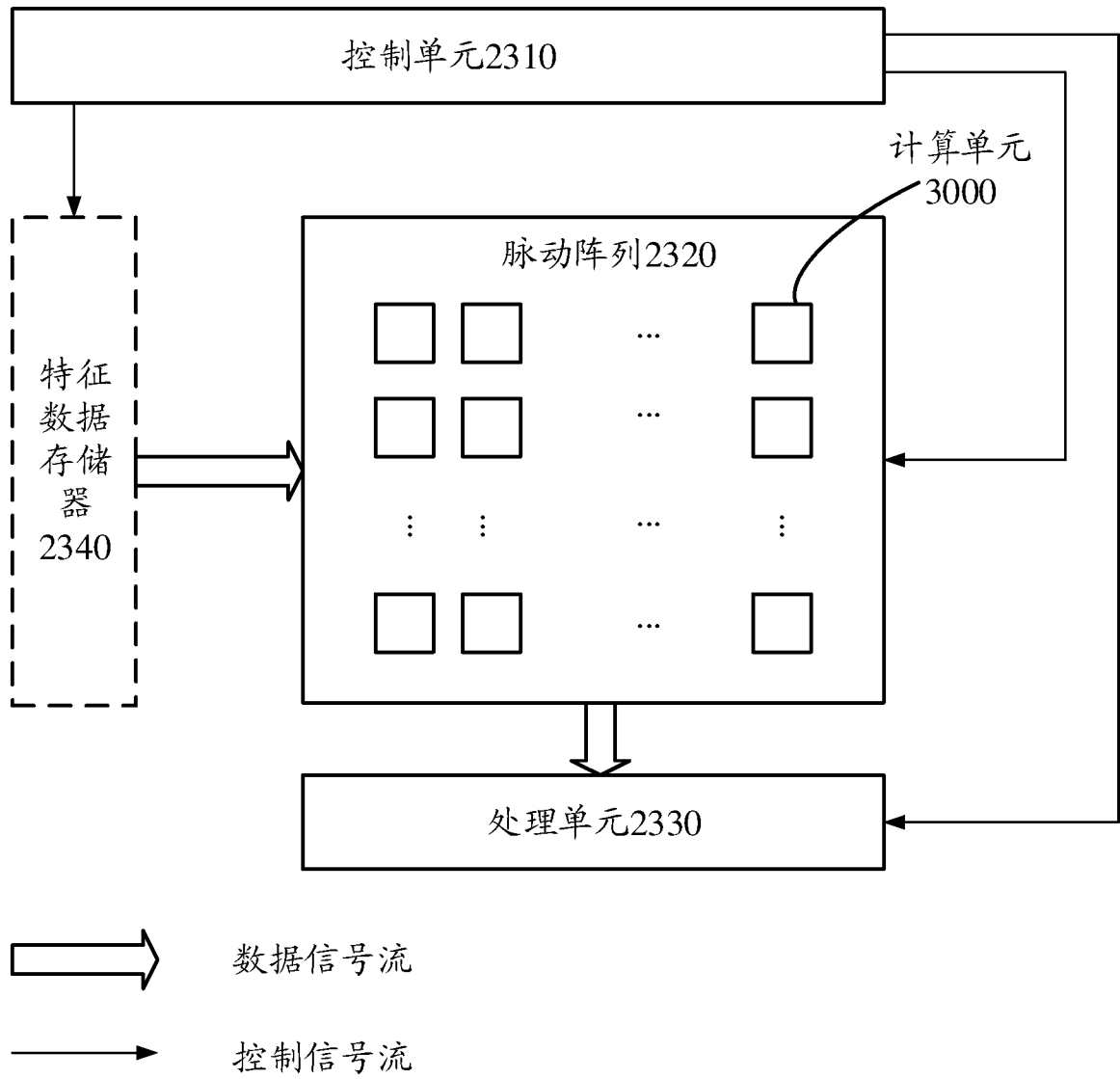


图 23

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2018/100335

A. CLASSIFICATION OF SUBJECT MATTER

G06N 3/06(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)

WPI, EPODOC, CNPAT, CNKI, IEEE, GOOGLE: 神经网络, 脉动, 阵列, 卷积, 池化, 计算, 中间, 结果, neural, network, pulsating, array, process+, unit?, element?, comput+, convolution, pool+, intermediate, result

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 107578095 A (INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES) 12 January 2018 (2018-01-12) claims 1-10, description, paragraphs [0007]-[0049], and figures 1-6	1-17
X	CN 107578098 A (INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES) 12 January 2018 (2018-01-12) claims 1-7, description, paragraphs [0007]-[0041], and figures 1-8	1-17
A	CN 107341544 A (TSINGHUA UNIVERSITY) 10 November 2017 (2017-11-10) entire document	1-17
A	CN 108182471 A (SHANGHAI YUEXIN ELECTRONIC TECHNOLOGY CO., LTD.) 19 June 2018 (2018-06-19) entire document	1-17

☐ 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

29 April 2019

Date of mailing of the international search report

15 May 2019

Name and mailing address of the ISA/CN

State Intellectual Property Office of the P. R. China
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/CN2018/100335

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	107578095	A	12 January 2018	CN 107578095 B	10 August 2018
CN	107578098	A	12 January 2018	None	
CN	107341544	A	10 November 2017	None	
CN	108182471	A	19 June 2018	None	

国际检索报告

国际申请号

PCT/CN2018/100335

A. 主题的分类 G06N 3/06 (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)) WPI, EPDOC, CNPAT, CNKI, IEEE, GOOGLE: 神经网络, 脉动, 阵列, 卷积, 池化, 计算, 中间, 结果, neural, network, pulsating, array, process+, unit?, element?, comput+, convolution, pool+, intermediate, result																	
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>X</td> <td>CN 107578095 A (中国科学院计算技术研究所) 2018年 1月 12日 (2018 - 01 - 12) 权利要求1-10, 说明书第[0007]-[0049]段, 附图1-6</td> <td>1-17</td> </tr> <tr> <td>X</td> <td>CN 107578098 A (中国科学院计算技术研究所) 2018年 1月 12日 (2018 - 01 - 12) 权利要求1-7, 说明书第[0007]-[0041]段, 附图1-8</td> <td>1-17</td> </tr> <tr> <td>A</td> <td>CN 107341544 A (清华大学) 2017年 11月 10日 (2017 - 11 - 10) 全文</td> <td>1-17</td> </tr> <tr> <td>A</td> <td>CN 108182471 A (上海岳芯电子科技有限公司) 2018年 6月 19日 (2018 - 06 - 19) 全文</td> <td>1-17</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	X	CN 107578095 A (中国科学院计算技术研究所) 2018年 1月 12日 (2018 - 01 - 12) 权利要求1-10, 说明书第[0007]-[0049]段, 附图1-6	1-17	X	CN 107578098 A (中国科学院计算技术研究所) 2018年 1月 12日 (2018 - 01 - 12) 权利要求1-7, 说明书第[0007]-[0041]段, 附图1-8	1-17	A	CN 107341544 A (清华大学) 2017年 11月 10日 (2017 - 11 - 10) 全文	1-17	A	CN 108182471 A (上海岳芯电子科技有限公司) 2018年 6月 19日 (2018 - 06 - 19) 全文	1-17
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求															
X	CN 107578095 A (中国科学院计算技术研究所) 2018年 1月 12日 (2018 - 01 - 12) 权利要求1-10, 说明书第[0007]-[0049]段, 附图1-6	1-17															
X	CN 107578098 A (中国科学院计算技术研究所) 2018年 1月 12日 (2018 - 01 - 12) 权利要求1-7, 说明书第[0007]-[0041]段, 附图1-8	1-17															
A	CN 107341544 A (清华大学) 2017年 11月 10日 (2017 - 11 - 10) 全文	1-17															
A	CN 108182471 A (上海岳芯电子科技有限公司) 2018年 6月 19日 (2018 - 06 - 19) 全文	1-17															
☐ 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 “&” 同族专利的文件																	
国际检索实际完成的日期 2019年 4月 29日		国际检索报告邮寄日期 2019年 5月 15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 李敏 电话号码 86-(10)-53961526															

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2018/100335

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	107578095	A	2018年 1月 12日	CN 107578095 B	2018年 8月 10日
CN	107578098	A	2018年 1月 12日	无	
CN	107341544	A	2017年 11月 10日	无	
CN	108182471	A	2018年 6月 19日	无	