

(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  
2023 年 9 月 21 日 (21.09.2023)



(10) 国际公布号  
**WO 2023/173530 A1**

(51) 国际专利分类号:  
**G11C 5/02** (2006.01)

(21) 国际申请号: PCT/CN2022/087794

(22) 国际申请日: 2022 年 4 月 20 日 (20.04.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:  
202210272801.8 2022 年 3 月 18 日 (18.03.2022) CN

(71) 申请人: 华中科技大学(HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY) [CN/CN]; 中国湖北省武汉洪山区珞喻路 1037 号, Hubei 430074 (CN)。

(72) 发明人: 缪向水(MIAO, Xiangshui); 中国湖北省武汉洪山区珞喻路 1037 号, Hubei 430074 (CN)。 付嘉伟(FU, Jiawei); 中国湖北省武汉洪山区珞喻路 1037 号, Hubei 430074 (CN)。 何毓辉

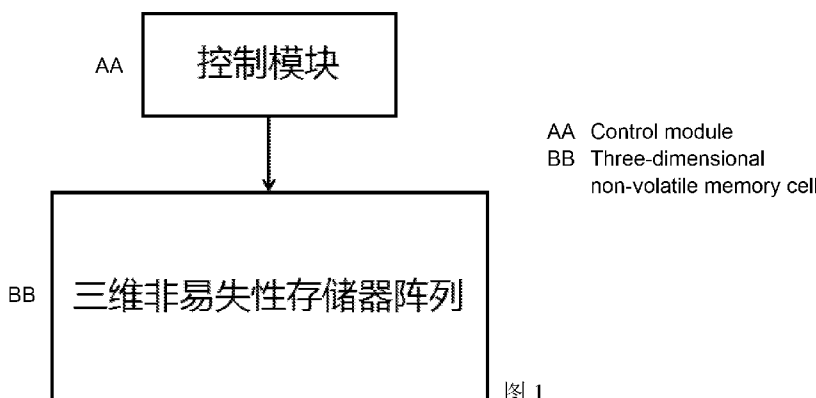
(HE, Yuhui); 中国湖北省武汉洪山区珞喻路 1037 号, Hubei 430074 (CN)。

(74) 代理人: 华中科技大学专利中心 (HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY PATENT AGENCY CENTER); 中国湖北省武汉洪山区珞喻路 1037 号华中科技大学东一楼 340 室, Hubei 430074 (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, IT, JM, 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, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(54) Title: CONVOLUTION OPERATION ACCELERATOR AND CONVOLUTION OPERATION METHOD

(54) 发明名称: 一种卷积运算加速器及卷积运算方法



(57) Abstract: A convolution operation accelerator and a convolution operation method, relating to the field of microelectronic devices. Each word line electrode connects one column of non-volatile memory cells in non-volatile memory cells arranged in a matrix, and bit line electrodes obliquely connect the non-volatile memory cells arranged in a matrix; input data of each word line can be subjected to multiply-add operation with upper-layer convolution kernel units and lower-layer convolution kernel units, to achieve natural sliding of the convolution kernel units in a y direction in two-dimensional input; the oblique bit lines and multiple copies of convolution kernels in each layer of non-volatile memory array can enable one piece of input data and convolution kernel data at different positions in a same convolution kernel unit to be subjected to multiplication operation, to achieve natural sliding of the convolution kernel units in an x direction in two-dimensional input. Convolution operation can be completed in parallel at one time without expanding the two-dimensional input into one dimension, the operation speed is high, and the degree of parallelism and the utilization rate of the array units are high.

(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, SC, 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)。

本国际公布:

— 包括国际检索报告 (条约第21条(3))。

---

(57) 摘要: 一种卷积运算加速器及卷积运算方法, 属于微电子器件领域; 其中, 每条字线电极将矩阵排列的非易失性存储单元中的一列连接起来, 每条位线电极倾斜地将矩阵排列的非易失性存储单元连接起来; 每条字线的输入数据可以和上下两层卷积核单元进行乘加操作, 实现卷积核单元在二维输入中y方向上的天然滑动; 倾斜的位线和卷积核在每层非易失性存储阵列中的多次复制可以使一个输入数据和相同卷积核单元中不同位置的卷积核数据进行乘法运算, 实现卷积核单元在二维输入中x方向上的天然滑动, 可以在不将二维输入展开为一维的情况下一次性并行完成卷积计算, 运算速度快, 并行程度及阵列单元的利用率均较高。

## 一种卷积运算加速器及卷积运算方法

### 【技术领域】

本发明属于微电子器件领域，更具体地，涉及一种卷积运算加速器及卷积运算方法。

### 【背景技术】

卷积神经网络(Convolution Neural network, CNN)是前馈神经网络的一种形式，通常包含卷积层、池化层和全连接层。卷积神经网络通常用来处理图片信息，将图片输入到卷积层后，经过卷积运算提取出局部特征，然后通过池化层压缩模型大小，减少计算量，最后通过全连接层将特征整合得到输出。卷积神经网络具有局部连接和权值共享的特点，其参数量远远小于多层感知机模型，在图像识别领域已经取得了良好的效果。

卷积神经网络虽然有较少的参数，但在前向传播过程中，卷积层中的卷积运算产生了网络中大部分的计算量，卷积运算过程为：卷积核在输入数据上以一定步长滑动，每滑动一次进行一次卷积核与对应输入数据的乘加运算，这在传统计算单元(例如 CPU、GPU)中计算速度慢、占用空间大。

同时对于一些基于新型存储器的计算结构，如平面交叉开关(crossbar)阵列来说，需要将二维的输入信息拉伸为一维然后输入给阵列；另外，如果要充分利用阵列单元，就必须采用滚动式输入，那么输入的并行程度很低，所需时间代价巨大；如果在 crossbar 阵列中复制多个卷积核进行大规模并行计算的话，会占用大量面积，同时大部分单元不会被使用，增加漏电流对计算结果的影响。

因此，需要一种能够高度并行计算并且能够充分利用计算资源的架构来加速卷积运算。

## 【发明内容】

针对现有技术的以上缺陷或改进需求，本发明提供了一种卷积运算加速器及卷积运算方法，用以解决平面存储器阵列进行卷积计算时需要将二维输入拉伸为一维、并行程度低以及阵列单元利用率低的问题。

为了实现上述目标，本发明提供了一种卷积运算加速器，包括：三维非易失性存储器阵列和控制模块；

三维非易失性存储器阵列包括： $P$  个字线电极层；任意相邻的两个字线电极层之间均放置有一个位线电极层；任意相邻的字线电极层和位线电极层之间均放置有一个非易失性存储单元阵列，且非易失性存储单元阵列与字线电极层和位线电极层均垂直相连；

字线电极层包括多个平行排布的字线电极； $P$  个字线电极层中的字线电极共同构成一个字线电极阵列；

非易失性存储单元阵列中的每一列非易失性存储单元均连接在与非易失性存储单元阵列相连的字线电极层中的同一条字线上；非易失性存储单元阵列中每一条斜线上的非易失性存储单元均连接在与非易失性存储单元阵列相连的位线电极阵列中的同一条位线上；其中，斜线为非易失性存储单元阵列中与非易失性存储单元阵列中的对应对角线平行的斜线；

记二维输入数据的尺寸为  $M \times N$ ；

当与二维输入数据进行卷积运算的卷积核尺寸为  $2k \times c$  时，控制模块用于将卷积核按行拆分为  $k$  个  $2 \times c$  大小的卷积核单元， $k$  为正整数；在字线电极阵列中选中  $k$  个大小为  $(M-2(k-1)) \times N$  的不同子阵列单元，并与  $k$  个  $2 \times c$  大小的卷积核单元按照卷积核的拆分顺序一一对应；将每个卷积核单元及其  $N-c$  次复制均存储至对应子阵列单元中所有相邻的两个字线电极层之间的两层非易失性存储单元阵列中；将二维输入数据上第  $2i-1$  行到第  $M-2(k-i)$  行的数据以电压的方式按照对应的坐标信息施加到第  $i$  个子阵列单元中的对应字线电极上， $i = 1, 2, \dots, k$ ；

当卷积核的尺寸为  $(2k+1) \times c$  时，控制模块用于将卷积核按行拆分为  $k$  个  $2 \times c$  大小的卷积核单元和一个  $1 \times c$  大小的卷积核单元；在字线电极阵列中选中  $k$  个大小为  $(M-2(k-1)-1) \times N$  的子阵列单元，以及一个大小为  $(M-2k) \times N$  的子阵列单元； $k$  个大小为  $(M-2(k-1)-1) \times N$  的子阵列单元与  $k$  个  $2 \times c$  大小的卷积核单元按照卷积核的拆分顺序一一对应；将每个  $2 \times c$  大小的卷积核单元及其  $N-c$  次复制均存储至对应子阵列单元中所有相邻的两个字线电极层之间的两层非易失性存储单元阵列中；对上述大小为  $(M-2k) \times N$  的子阵列单元中的各字线电极层，在与其相连的其中一个非易失性存储单元阵列中分别存储  $1 \times c$  大小的卷积核单元及该  $1 \times c$  大小的卷积核单元的  $N-c$  次复制，未存储卷积核单元的非易失性存储单元阵列中的非易失存储单元全部置为高阻态；将二维输入数据上第  $2i-1$  行到第  $M-2(k-i)-1$  行的数据以电压的方式按照对应的坐标信息施加到第  $i$  个  $(M-2(k-1)-1) \times N$  大小的子阵列单元中的对应字线电极上， $i = 1, 2, \dots, k$ ；将二维输入数据上第  $2k+1$  行到第  $M$  行的数据以电压的方式按照对应的坐标信息施加到上述大小为  $(M-2k) \times N$  的子阵列单元中的对应字线电极上；

三维非易失性存储器阵列用于基于非易失性存储单元阵列并行地实现卷积核单元与二维输入数据不同部分的点乘运算，经由对应的位线电极层并行输出卷积核单元与二维输入数据对应部分的点乘运算结果之和，从而实现卷积核与二维输入数据的卷积运算；

其中，卷积核单元的每一行卷积核数据在对应非易失性存储单元阵列的对应斜线上依次存储后，沿着卷积核滑动方向水平移动到相邻的  $N-c$  条斜线上再次对该行卷积核数据进行存储，实现每一层非易失性存储单元阵列存储卷积核单元中对应行的卷积核数据及该卷积核数据的  $N-c$  次复制。

进一步优选地，当卷积核的尺寸为  $2k \times c$  时，三维非易失性存储器阵列用于基于上述大小为  $(M-2(k-1)) \times N$  的子阵列单元中所有相邻的两个字线电极层之间的两层非易失性存储单元阵列，并行地实现  $2 \times c$  大小的卷积核单

元与二维输入数据不同部分的点乘运算后, 经由对应的位线电极层并行输出  $2 \times c$  大小的卷积核单元与二维输入数据对应部分的点乘运算结果之和; 控制模块还用于对所选中的各子阵列单元中同一相对位置的位线输出分别进行汇总, 得到卷积核与二维输入数据的卷积运算结果;

当卷积核的尺寸为  $(2k+1) \times c$  时, 三维非易失性存储器阵列用于基于上述大小为  $(M-2(k-1)-1) \times N$  的子阵列单元中所有相邻的两个字线电极层之间的两层非易失性存储单元阵列, 并行地实现  $2 \times c$  大小的卷积核单元与二维输入数据不同部分的点乘运算后, 经由对应的位线电极层并行输出  $2 \times c$  大小的卷积核单元与二维输入数据对应部分的点乘运算结果之和; 同时基于上述大小为  $(M-2k) \times N$  的子阵列单元中的非易失性存储单元阵列, 并行地实现  $1 \times c$  大小的卷积核单元与二维输入数据不同部分的点乘运算后, 经由对应的位线电极层并行输出  $1 \times c$  大小的卷积核单元与二维输入数据对应部分的点乘运算结果之和; 控制模块还用于将所选中的各子阵列单元中的各位线输出对应进行汇总, 得到卷积核与二维输入数据的卷积运算结果。

进一步优选地, 三维非易失性存储器阵列为多层堆叠结构, 由字线电极层、非易失性存储单元阵列和位线电极层依次堆叠形成; 各字线电极层之间、字线电极层与位线电极层之间均互相平行。

进一步优选地, 控制模块包括: 输入端信号发生装置、输出端信号发生装置和输出信号读出装置;

输入端信号发生装置和输出端信号发生装置相互配合, 在子阵列单元所对应的字线电极和位线电极上施加电压, 选择对应的非易失存储单元, 并改变所选中的各非易失存储单元的电导值, 使其存储对应的卷积核数据;

输入端信号发生装置还用于将二维输入数据以电压的方式, 按照二维输入数据的坐标信息施加到上述子阵列单元中的对应字线电极上;

输出信号读出装置用于检测位线电极上的输出电流, 并转化为电压信号, 作为卷积运算的结果。

进一步优选地，输入信号发生装置的读电压小于或等于非易失存储单元上的当前 set 电压值，以保证在进行卷积运算时非易失存储单元的电导值不被改变。

进一步优选地，上述卷积运算加速器用于实现卷积神经网络中的卷积运算。

进一步优选地，在卷积神经网络的训练过程中，上述控制模块还用于通过在对应的字线电极和位线电极上施加写电压来调整各非易失存储单元的电导值，以对卷积神经网络的参数进行更新，并将卷积运算加速器上一轮的卷积运算结果重新作为卷积运算加速器下一轮卷积运算的输入。

进一步优选地，当需要增加非易失存储单元的电导值时，在对应的字线电极和位线电极上施加的写电压差值的绝对值为对应的 set 电压值；

当需要减小非易失存储单元的电导值时，在对应的字线电极和位线电极上施加的写电压差值的绝对值为对应的 reset 电压值。

第二方面，本发明提供了一种基于上述卷积运算加速器的卷积运算方法，包括以下步骤：

S1、当与二维输入数据进行卷积运算的卷积核尺寸为  $2k \times c$  时，将卷积核按行拆分为  $k$  个  $2 \times c$  大小的卷积核单元， $k$  为正整数；在字线电极阵列中选中  $k$  个大小为  $(M-2(k-1)) \times N$  的不同子阵列单元，并与  $k$  个  $2 \times c$  大小的卷积核单元按照卷积核的拆分顺序一一对应；将每个卷积核单元及其  $N-c$  次复制均存储至对应子阵列单元中所有相邻的两个字线电极层之间的两层非易失性存储单元阵列中；将二维输入数据上第  $2i-1$  行到第  $M-2(k-i)$  行的数据以电压的方式按照对应的坐标信息施加到第  $i$  个子阵列单元中的对应字线电极上， $i = 1, 2, \dots, k$ ；

当卷积核的尺寸为  $(2k+1) \times c$  时，将卷积核按行拆分为  $k$  个  $2 \times c$  大小的卷积核单元和一个  $1 \times c$  大小的卷积核单元；在字线电极阵列中选中  $k$  个大小为  $(M-2(k-1)-1) \times N$  的子阵列单元，以及一个大小为  $(M-2k) \times N$  的子阵

列单元； $k$  个大小为  $(M-2(k-1)-1) \times N$  的子阵列单元与  $k$  个  $2 \times c$  大小的卷积核单元按照卷积核的拆分顺序一一对应；将每个  $2 \times c$  大小的卷积核单元及其  $N-c$  次复制均存储至对应子阵列单元中所有相邻的两个字线电极层之间的两层非易失性存储单元阵列中；对上述大小为  $(M-2k) \times N$  的子阵列单元中的各字线电极层，在与其相连的其中一个非易失性存储单元阵列中分别存储  $1 \times c$  大小的卷积核单元及该  $1 \times c$  大小的卷积核单元的  $N-c$  次复制，未存储卷积核单元的非易失性存储单元阵列中的非易失存储单元全部置为高阻态；将二维输入数据上第  $2i-1$  行到第  $M-2(k-i)-1$  行的数据以电压的方式按照对应的坐标信息施加到第  $i$  个  $(M-2(k-1)-1) \times N$  大小的子阵列单元中的对应字线电极上， $i = 1, 2, \dots, k$ ；将二维输入数据上第  $2k+1$  行到第  $M$  行的数据以电压的方式按照对应的坐标信息施加到上述大小为  $(M-2k) \times N$  的子阵列单元中的对应字线电极上；

S2、基于非易失性存储单元阵列并行地实现卷积核单元与二维输入数据不同部分的点乘运算，经由对应的位线电极层并行输出卷积核单元与二维输入数据对应部分的点乘运算结果之和，从而实现卷积核与二维输入数据的卷积运算；

其中，二维输入数据的尺寸为  $M \times N$ ；卷积核单元的每一行卷积核数据在对应非易失性存储单元阵列的对应斜线上依次存储后，沿着卷积核滑动方向水平移动到相邻的  $N-c$  条斜线上再次对该行卷积核数据进行存储，实现每一层非易失性存储单元阵列存储卷积核单元中对应行的卷积核数据及该卷积核数据的  $N-c$  次复制。

进一步优选地，当卷积核的尺寸为  $2k \times c$  时，步骤 S2 包括：基于上述大小为  $(M-2(k-1)) \times N$  的子阵列单元中所有相邻的两个字线电极层之间的两层非易失性存储单元阵列，并行地实现  $2 \times c$  大小的卷积核单元与二维输入数据不同部分的点乘运算后，经由对应的位线电极层并行输出  $2 \times c$  大小的卷积核单元与二维输入数据对应部分的点乘运算结果之和，并对所选中的

各子阵列单元中同一相对位置的位线输出分别进行汇总，得到卷积核与二维输入数据的卷积运算结果；

当卷积核的尺寸为  $(2k+1) \times c$  时，步骤 S2 包括：基于上述大小为  $(M-2(k-1)-1) \times N$  的子阵列单元中所有相邻的两个字线电极层之间的两层非易失性存储单元阵列，并行地实现  $2 \times c$  大小的卷积核单元与二维输入数据不同部分的点乘运算后，经由对应的位线电极层并行输出  $2 \times c$  大小的卷积核单元与二维输入数据对应部分的点乘运算结果之和；同时基于上述大小为  $(M-2k) \times N$  的子阵列单元中的非易失性存储单元阵列，并行地实现  $1 \times c$  大小的卷积核单元与二维输入数据不同部分的点乘运算后，经由对应的位线电极层并行输出  $1 \times c$  大小的卷积核单元与二维输入数据对应部分的点乘运算结果之和；并将所选中的各子阵列单元中的各位线输出对应进行汇总，得到卷积核与二维输入数据的卷积运算结果。

总体而言，通过本发明所构思的以上技术方案，能够取得以下有益效果：

1、本发明提供了一种卷积运算加速器，其中，每条字线电极将矩阵排列的非易失性存储单元中的一列连接起来，每条位线电极倾斜地将矩阵排列的非易失性存储单元连接起来；每条字线的输入数据可以和上下两层卷积核单元进行乘加操作，实现卷积核单元在二维输入中  $y$  方向上的天然滑动；倾斜的位线和卷积核在每层非易失性存储阵列中的多次复制可以使一个输入数据和相同卷积核单元中不同位置的卷积核数据进行乘法运算，实现卷积核单元在二维输入中  $x$  方向上的天然滑动，因此可以在不将二维输入展开为一维的情况下一次性并行完成卷积计算，数据复用率和运算的并行程度均较高、运算速度快；与传统平面阵列相比，本发明利用三维阵列的高集成度特性，在较小的面积下容纳更多单元，执行更多次的计算，大大提高了阵列单元的利用率。

2、本发明所提供的卷积运算加速器，分别考虑卷积核行为奇数和偶数的情况进行拆分并存储，能够并行地实现不同拆分部分的与对应二维数据

的卷积运算，充分利用了三维非易失性存储器阵列的计算资源，运算速度较快。

3、本发明所提供的卷积运算加速器可以用来实现卷积神经网络的功能，将神经网络的输入、权重和输出映射为三维非易失性存储器的输入电压、电导以及输出电流，同时由于三维非易失性存储器阵列拓扑结构的独特设计，阵列层面的物理计算过程能够天然匹配卷积神经网络中大量存在的卷积核滑动运算过程，能够在保留二维数据的位置信息的情况下一一次性完成卷积运算，并且获得二维输出数据，高度还原了软件中的卷积神经网络的计算过程，同时利用尽可能少的输入输出端口实现卷积高度并行计算，大大提高了资源利用率和运算速度。

4、本发明基于三维非易失性存储器阵列的卷积运算，相比于传统的冯诺依曼架构计算单元例如 CPU、GPU 等，没有存储器和计算单元之间的数据传输速度限制，实现了存算一体，能够用于实现高效大规模并行的卷积神经网络计算。

## 【附图说明】

图 1 为本发明第一方面提供的一种卷积运算加速器的结构示意图；

图 2 为本发明实施例 1 提供的卷积运算加速器的结构示意图；

图 3 为本发明实施例 1 提供的三维非易失性存储器阵列的结构示意图；

图 4 为本发明实施例 1 提供的三维非易失性存储器的俯视剖面图；

图 5 为本发明实施例 2 提供的基于实施例 1 所述三维非易失性存储器阵列的卷积运算的流程图；

图 6 为本发明实施例 2 提供的三维非易失性存储器阵列天然匹配卷积神经网络中卷积核在输入数据的 x 方向上滑动执行乘加运算的示意图；

图 7 为本发明实施例 2 提供的三维非易失性存储阵列天然匹配卷积神经网络中卷积核在输入数据的 y 方向上滑动执行乘加运算的示意图；

图 8 为本发明实施例 3 提供的横向上具有八条字线的三维非易失性存储阵列的主视图。

图 9 为本发明实施例 3 提供的横向上具有八条字线的三维非易失性存储阵列的俯视剖视图。

### 【具体实施方式】

为了使本发明的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本发明进行进一步详细说明。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。此外，下面所描述的本发明各个实施方式中所涉及到的技术特征只要彼此之间未构成冲突就可以相互组合。

为了实现上述目的，如图 1 所示，本发明提供了一种卷积运算加速器，包括：三维非易失性存储器阵列和控制模块；

其中，三维非易失性存储器阵列包括： $P$  个字线电极层；任意相邻的两个字线电极层之间均放置有一个位线电极层；任意相邻的字线电极层和位线电极层之间均放置有一个非易失性存储单元阵列，且非易失性存储单元阵列与字线电极层和位线电极层均垂直相连；具体地，三维非易失性存储器阵列为多层堆叠结构，由字线电极层、非易失性存储单元阵列和位线电极层依次堆叠形成；各字线电极层之间、字线电极层与位线电极层之间均互相平行；

字线电极层包括多个平行排布的字线电极； $P$  个字线电极层中的字线电极共同构成一个字线电极阵列；

非易失性存储单元阵列中的每一列非易失性存储单元均连接在与非易失性存储单元阵列相连的字线电极层中的同一条字线上；非易失性存储单元阵列中每一条斜线上的非易失性存储单元均连接在与非易失性存储单元阵列相连的位线电极阵列中的同一条位线上；其中，斜线为非易失性存储单

元阵列中与非易失性存储单元阵列中的对应对角线平行的斜线;

记二维输入数据的尺寸为  $M \times N$ ;

当与二维输入数据进行卷积运算的卷积核尺寸为  $2k \times c$  时, 控制模块用于将卷积核按行拆分为  $k$  个  $2 \times c$  大小的卷积核单元,  $k$  为正整数; 在字线电极阵列中选中  $k$  个大小为  $(M-2(k-1)) \times N$  的不同子阵列单元, 并与  $k$  个  $2 \times c$  大小的卷积核单元按照卷积核的拆分顺序一一对应; 将每个卷积核单元及其  $N-c$  次复制均存储至对应子阵列单元中所有相邻的两个字线电极层之间的两层非易失性存储单元阵列中; 将二维输入数据上第  $2i-1$  行到第  $M-2(k-i)$  行的数据以电压的方式按照对应的坐标信息施加到第  $i$  个子阵列单元中的对应字线电极上,  $i = 1, 2, \dots, k$ ;

当卷积核的尺寸为  $(2k+1) \times c$  时, 控制模块用于将卷积核按行拆分为  $k$  个  $2 \times c$  大小的卷积核单元和一个  $1 \times c$  大小的卷积核单元; 在字线电极阵列中选中  $k$  个大小为  $(M-2(k-1)-1) \times N$  的子阵列单元, 以及一个大小为  $(M-2k) \times N$  的子阵列单元;  $k$  个大小为  $(M-2(k-1)-1) \times N$  的子阵列单元与  $k$  个  $2 \times c$  大小的卷积核单元按照卷积核的拆分顺序一一对应; 将每个  $2 \times c$  大小的卷积核单元及其  $N-c$  次复制均存储至对应子阵列单元中所有相邻的两个字线电极层之间的两层非易失性存储单元阵列中; 对上述大小为  $(M-2k) \times N$  的子阵列单元中的各字线电极层, 在与其相连的其中一个非易失性存储单元阵列中分别存储  $1 \times c$  大小的卷积核单元及该  $1 \times c$  大小的卷积核单元的  $N-c$  次复制, 未存储卷积核单元的非易失性存储单元阵列中的非易失存储单元全部置为高阻态; 将二维输入数据上第  $2i-1$  行到第  $M-2(k-i)-1$  行的数据以电压的方式按照对应的坐标信息施加到第  $i$  个  $(M-2(k-1)-1) \times N$  大小的子阵列单元中的对应字线电极上,  $i = 1, 2, \dots, k$ ; 将二维输入数据上第  $2k+1$  行到第  $M$  行的数据以电压的方式按照对应的坐标信息施加到上述大小为  $(M-2k) \times N$  的子阵列单元中的对应字线电极上;

三维非易失性存储器阵列用于基于非易失性存储单元阵列并行地实现卷积核单元与二维输入数据不同部分的点乘运算，经由对应的位线电极层并行输出卷积核单元与二维输入数据对应部分的点乘运算结果之和，从而实现卷积核与二维输入数据的卷积运算；

其中，卷积核单元的每一行卷积核数据在对应非易失性存储单元阵列的对应斜线上依次存储后，沿着卷积核滑动方向水平移动到相邻的  $N-c$  条斜线上再次对该行卷积核数据进行存储，实现每一层非易失性存储单元阵列存储卷积核单元中对应行的卷积核数据及该卷积核数据的  $N-c$  次复制。

具体地，当卷积核的尺寸为  $2k \times c$  时，三维非易失性存储器阵列用于基于上述大小为  $(M-2(k-1)) \times N$  的子阵列单元中所有相邻的两个字线电极层之间的两层非易失性存储单元阵列，并行地实现  $2 \times c$  大小的卷积核单元与二维输入数据不同部分的点乘运算后，经由对应的位线电极层并行输出  $2 \times c$  大小的卷积核单元与二维输入数据对应部分的点乘运算结果之和；控制模块还用于对所选中的各子阵列单元中同一相对位置的位线输出分别进行汇总，得到卷积核与二维输入数据的卷积运算结果；

当卷积核的尺寸为  $(2k+1) \times c$  时，三维非易失性存储器阵列用于基于上述大小为  $(M-2(k-1)-1) \times N$  的子阵列单元中所有相邻的两个字线电极层之间的两层非易失性存储单元阵列，并行地实现  $2 \times c$  大小的卷积核单元与二维输入数据不同部分的点乘运算后，经由对应的位线电极层并行输出  $2 \times c$  大小的卷积核单元与二维输入数据对应部分的点乘运算结果之和；同时基于上述大小为  $(M-2k) \times N$  的子阵列单元中的非易失性存储单元阵列，并行地实现  $1 \times c$  大小的卷积核单元与二维输入数据不同部分的点乘运算后，经由对应的位线电极层并行输出  $1 \times c$  大小的卷积核单元与二维输入数据对应部分的点乘运算结果之和；控制模块还用于将所选中的各子阵列单元中的各位线输出对应进行汇总，得到卷积核与二维输入数据的卷积运算结果。其中，各子阵列单元的输出均为  $(M-2k) \times (N-c)$ ，对应加和后即可以得到最终

的卷积运算结果。

在一种可选实施方式下，当卷积核的尺寸为  $2k \times c$  时，纵向上至少选中  $M-2(k-1)$  层字线层，横向上至少选中  $k \times N$  条字线；控制模块将卷积核拆分为  $k$  个  $2 \times c$  大小的卷积核单元，将横向上有  $k \times N$  个字线上的三维非易失性存储阵列视为  $k$  个具有  $N$  个字线的子阵列单元，在每一个子阵列单元的对应位置映射一个  $2 \times c$  大小的卷积核单元，二维输入数据在  $k$  个阵列单元上同样也会滑动，例如， $M \times N$  个输入数据，第一个阵列单元字线上映射第一行到第  $M-2(k-1)$  行数据，第二个阵列单元字线上映射第  $2+1$  行到  $M-2(k-2)$  行数据，第  $i$  个阵列单元字线上映射第  $2i-1$  行到第  $M-2(k-i)$  ( $i \in 1, 2, 3, \dots, k$ ) 行数据。此时，控制模块还用于对  $k$  个子阵列单元的中同一相对位置的位线输出分别进行汇总，得到最终的卷积运算结果。

当卷积核的尺寸为  $(2k+1) \times c$  时， $k$  个  $2 \times c$  大小的卷积核的映射方式和卷积核的尺寸为  $2k \times c$  时的方法类似，区别在于，在横向上选中  $(k+1) \times N$  个字线的三维非易失性存储阵列，将横向上有  $(k+1) \times N$  个字线上的三维非易失性存储阵列视为  $k+1$  个具有  $N$  个字线的子阵列单元，第  $i$  个阵列单元字线上映射第  $2i-1$  行到第  $M-2(k-i)-1$  ( $i \in 1, 2, 3, \dots, k$ ) 行数据；对于剩余的一个  $1 \times c$  大小的卷积核单元，将其映射到第  $k+1$  个阵列单元上，对于每层位线连接的两层非易失性存储单元，将其上方(或下方)的非易失性存储单元阵列中存储  $1 \times c$  大小的卷积核单元及其  $N-c$  次复制，将位线下方(或上方)的非易失性存储单元全部置为高阻态（视为断路）；本具体实施方式在纵向上选中  $M-2(k-1)-1$  层字线层，横向上选中  $(k+1) \times N$  条字线；当在上方的非易失性存储单元阵列上存储  $1 \times c$  大小的卷积核单元时，在从上至下数第 1 层到第  $M-2k$  层上映射二维输入数据中的第  $2k+1$  行到第  $M$  行数据。当在下方的非易失性存储单元阵列上存储  $1 \times c$  大小的卷积核单元时，在从上至下数第 2 层到第  $M-2k+1$  层上映射二维输入数据中的第  $2k+1$  行到第  $M$  行数据。此时，控制模块还用于对  $k+1$  个子阵列单元的中同一相对位置的位线输出分别进行

汇总，得到最终的卷积运算结果。

为了进一步说明本发明所提供的卷积运算加速器，下面结合具体实施例进行详述：

#### 实施例 1、

本实施例中的卷积运算加速器如图 2 所示，包括：三维非易失性存储器阵列和控制模块，其中，控制模块包括：输入端信号发生装置、输出端信号发生装置和输出信号读出装置；

如图 3 所示为三维非易失性存储器阵列的结构示意图，包括字线电极 1、位线电极 2、非易失存储单元 3。输入端信号发生装置和字线电极相连，输出端信号发生装置和输出信号读出装置连接至位线电极。V11~V44 以电压形式施加到字线电极上作为输入，I11~I33 为输入电压作用到非易失性存储阵列后得到的输出电流。

为了进一步说明三维非易失性存储器阵列内部的连接结构，如图 4 展示了三维非易失性存储器的俯视剖面图。左边所示的俯视剖面图为字线电极和非易失性存储单元阵列的两层结构，非易失性存储单元阵列在字线电极上方。字线 1~4 为平行排布的电极，每条字线上各自均匀分布着非易失性存储单元，非易失性存储单元 411~441 连接至字线 1，非易失性存储单元 412~442 连接至字线 2，非易失性存储单元 413~443 连接至字线 3，非易失性存储单元 414~444 连接至字线 4。右边所示的俯视剖面图在左边的俯视剖面图基础上添加了一层位线电极，自下而上分别为字线电极、非易失性存储器阵列和位线电极。位线电极平行排布，但与字线电极的方向呈一定角度。非易失性存储单元 411 连接至位线 1，非易失性存储单元 421、412 连接至位线 2，非易失性存储单元 431、422、413 连接至位线 3，非易失性存储单元 441、432、423、414 连接至位线 4。

需要说明的是，在三维非易失性存储器阵列中，每条字线将矩阵排列的非易失性存储单元中的一列连接起来，每条位线倾斜地将矩阵排列的非易

失性存储单元连接起来。每条字线的输入数据可以和上下两层卷积核单元进行乘加操作，实现卷积核单元在二维输入中  $y$  方向上的天然滑动。倾斜的位线和卷积核在每层非易失性存储阵列中的多次复制可以使一个输入数据和相同卷积核中不同位置的卷积核数据进行乘法运算，实现卷积核在二维输入中  $x$  方向上的天然滑动，因此可以在不将二维输入展开为一维的情况下一次性并行完成卷积计算，运算速度极快。

进一步地，输入端信号发生装置和输出端信号发生装置相互配合，在子阵列单元所对应的字线电极和位线电极上施加电压，选择对应的非易失存储单元，并改变所选中的各非易失存储单元的电导值，使其存储对应的卷积核数据；

输入端信号发生装置还用于将二维输入数据以电压的方式，按照二维输入数据的坐标信息施加到上述子阵列单元中的对应字线电极上；

输出信号读出装置用于检测位线电极上的输出电流，并转化为电压信号，作为卷积运算的结果。

具体地，在卷积运算过程中，输入端信号发生装置用于将输入数据对应的电压施加在对应的字线电极上，且此时电压幅值不超过非易失存储单元的  $set$  电压。输出信号读出装置用于将位线电极上的输出电流转化为电压，并读出此电压值，作为卷积运算的结果。

需要说明的是，上述卷积运算加速器用于实现卷积神经网络中的卷积运算。此时，三维非易失性存储器阵列的字线电极作为卷积神经网络的输入，每层字线电极对应卷积神经网络二维输入中的一行，每层非易失存储阵列对应卷积核中的一行，每层位线电极对应着神经网络二维输出中的一行。输入端信号发生装置用于在字线电极上施加电压，所施加的电压有两个用途，一是用电压表示输入数据，经过调整好的电导值得到电流输出，完成神经网络的前向传播运算；二在字线电极上施加电压并配合位线电极来选中一个非易失性存储单元，并改变选中的非易失存储单元的电导值。输出端信号发

生装置用于在位线电极上施加电压，配合字线电极选中一个非易失性存储单元，并改变选中的非易失性存储单元的电导值。输出信号读出装置用于检测位线电极上的输出电流，并转化为电压信号，作为卷积运算的结果。另外，在卷积神经网络的训练过程中，上述控制模块还用于通过在对应的字线电极和位线电极上施加写电压来调整各非易失存储单元的电导值，以对卷积神经网络的参数进行更新，并将卷积运算加速器上一轮的卷积运算结果重新作为卷积运算加速器下一轮卷积运算的输入。具体地，输入端信号发生装置用于在与需要进行电导调节的非易失性存储单元连接的字线电极上施加一定幅值的电压，同时输出端信号发生装置在与需要进行电导调节的非易失性存储单元连接的位线电极上施加一定幅值的电压，当一个非易失性存储单元两端的字线电极和位线电极都被施加了电压后，则代表此单元被选中，而输入端信号发生装置或输出端信号发生装置单独施加的电压不能改变非易失存储单元的电导值；输入端信号发生装置和输出端信号发生装置施加的写电压相减之后的绝对值需等于 set 电压值（置态电压）或者 reset 电压值（重置电压），取决于在更新过程中电导值需要被增加或者减少；此时，单元两端电压差等于此单元的 set 电压或者 reset 电压；具体地，需要增加电导时，非易失性存储单元两侧电压等于 set 电压；需要减少电导时，非易失性存储单元两侧电压等于 reset 电压，直至调节为卷积核权重对应的电导值。

第二方面，本发明提供了基于上述卷积运算加速器的卷积运算方法，包括以下步骤：

S1、当与二维输入数据进行卷积运算的卷积核尺寸为  $2k \times c$  时，将卷积核按行拆分为  $k$  个  $2 \times c$  大小的卷积核单元， $k$  为正整数；在字线电极阵列中选中  $k$  个大小为  $(M-2(k-1)) \times N$  的不同子阵列单元，并与  $k$  个  $2 \times c$  大小的卷积核单元按照卷积核的拆分顺序一一对应；将每个卷积核单元及其  $N-c$  次复制均存储至对应子阵列单元中所有相邻的两个字线电极层之间的两层非

易失性存储单元阵列中；将二维输入数据上第  $2i-1$  行到第  $M-2(k-i)$  行的数据以电压的方式按照对应的坐标信息施加到第  $i$  个子阵列单元中的对应字线电极上， $i = 1, 2, \dots, k$ ；

当卷积核的尺寸为  $(2k+1) \times c$  时，将卷积核按行拆分为  $k$  个  $2 \times c$  大小的卷积核单元和一个  $1 \times c$  大小的卷积核单元；在字线电极阵列中选中  $k$  个大小为  $(M-2(k-1)-1) \times N$  的子阵列单元，以及一个大小为  $(M-2k) \times N$  的子阵列单元； $k$  个大小为  $(M-2(k-1)-1) \times N$  的子阵列单元与  $k$  个  $2 \times c$  大小的卷积核单元按照卷积核的拆分顺序一一对应；将每个  $2 \times c$  大小的卷积核单元及其  $N-c$  次复制均存储至对应子阵列单元中所有相邻的两个字线电极层之间的两层非易失性存储单元阵列中；对上述大小为  $(M-2k) \times N$  的子阵列单元中的各字线电极层，在与其相连的其中一个非易失性存储单元阵列中分别存储  $1 \times c$  大小的卷积核单元及该  $1 \times c$  大小的卷积核单元的  $N-c$  次复制，未存储卷积核单元的非易失性存储单元阵列中的非易失存储单元全部置为高阻态；将二维输入数据上第  $2i-1$  行到第  $M-2(k-i)-1$  行的数据以电压的方式按照对应的坐标信息施加到第  $i$  个  $(M-2(k-1)-1) \times N$  大小的子阵列单元中的对应字线电极上， $i = 1, 2, \dots, k$ ；将二维输入数据上第  $2k+1$  行到第  $M$  行的数据以电压的方式按照对应的坐标信息施加到上述大小为  $(M-2k) \times N$  的子阵列单元中的对应字线电极上；

具体地，通过在字线和位线选中非易失存储单元，在字线和位线上施加电压增加或者减少选中的非易失存储单元的电导值；在上述过程中，比较被选中单元的电导值和目标电导值，若不相等，则通过字线和位线在该单元两端施加电压脉冲使其电导值和目标电导值相等，最终目标为将非易失存储单元的电导值调节为对应的卷积核数据。

进一步地，在输入二维数据时，将二维输入数据转化为对应电压值，并通过输入信号发生装置施加到对应的字线电极上；其中，二维排布的输入数

据按照坐标信息一一对应到二维排布的字线电极上。

S2、基于非易失性存储单元阵列并行地实现卷积核单元与二维输入数据不同部分的点乘运算，经由对应的位线电极层并行输出卷积核单元与二维输入数据对应部分的点乘运算结果之和，从而实现卷积核与二维输入数据的卷积运算；

其中，二维输入数据的尺寸为  $M \times N$ ；卷积核单元的每一行卷积核数据在对应非易失性存储单元阵列的对应斜线上依次存储后，沿着卷积核滑动方向水平移动到相邻的  $N-c$  条斜线上再次对该行卷积核数据进行存储，实现每一层非易失性存储单元阵列存储卷积核单元中对应行的卷积核数据及该卷积核数据的  $N-c$  次复制。

具体地，施加在字线电极上的电压作用在非易失性存储阵列后，在位线电极上产生电流，使用输出信号读出装置将位线上的电流转化为电压，得到输出电压矩阵。

为了进一步说明本发明所提供的卷积运算方法，下面结合具体实施例进行详述：

#### 实施例 2、

为进一步说明本发明提出的基于三维非易失性存储器阵列的卷积运算方法以及其优势，下面以  $4 \times 4$  大小的图片作为输入信息， $2 \times 2$  大小的卷积核为例，阐述本发明提出的卷积运算方法的技术细节。

图 5 展示了基于实施例 1 所述三维非易失性存储器阵列的卷积运算的流程图，包含以下步骤：

A1、将  $4 \times 4$  大小的图片中共 16 个像素值映射为电压值，并确定映射后的电压值对应的字线电极；

具体的，将 0~255 范围内的像素值线性映射到一定范围内的电压值，此电压范围根据不同类型的非易失存储器选择，遵循最大电压值不超过非易失存储器的 set 电压原则。设非易失性存储器的 set 电压为  $V_{set}$ ，则像素值

对应的电压值为  $\frac{P}{255} \times V_{set}$ ，其中  $P$  为像素值，255 为像素值的最大值，例如  $V_{set}$  为 1V，则映射的电压范围在 0~1V，当像素值为 127.5 时，此像素值映射后的电压值为  $\frac{127.5}{255} \times 1 = 0.5V$ 。

具体的，4×4 图片中的第一行第一列的像素值映射的电压值对应图 3 中的 V11，在 V11 对应的字线电极中输入，第二行第一列的像素值映射的电压值对应图 3 中的 V12，依此类推，则图 5 中 V11~V44 构成的 4×4 电压矩阵一一对应着 4×4 图片中的像素点。

A2、将 2×2 的卷积核权重映射到非易失性存储阵列上，即分配卷积核权重在非易失性存储阵列上的位置；

具体的，以图 6 中一层非易失性存储器阵列的映射方式为例，V41~V44 对输入图片数据的第四行数据对应的电压值，k11~k22 为 2×2 卷积核的权重，其中 k11、k12 为卷积核中第一行的权重，k21、k22 为卷积核中第二行的权重，按照卷积运算规则，卷积核中第二行的权重需要在第四行的输入数据上滑动进行乘加运算，对应着 k21、k22 需要在 V41~V44 上滑动三次，于是将 k21 和 k22 复制三次映射到图 5 中左上图所示的位置，分别对应着图 4 中 421 和 412、422 和 413、423 和 414 位置，其他非易失性存储器层的卷积核映射方式可以依次类推。

图 7 为三维非易失性存储器的侧视图，展示了卷积核在其他非易失存储单元阵列层上的映射方案，其中 V11~V41 为输入图片数据的第一列数据对应的电压值，以 V21 为例，在卷积运算中卷积核的滑动过程中，需要分别和 k21、k11 进行乘法运算，所以 V21 对应的字线电极相连的上下两个非易失性存储单元分别映射 k21 和 k11，同时每个非易失性存储阵列层内的卷积核映射方式可以结合图 6 以及上段所述内容类推。所以，每两层非易失存储单元阵列完整映射一个 2×2 的卷积核的三次复制，而在 y 方向上卷积核同样需要滑动三次，所以需要六层非易失存储器。

A3、根据卷积核权重值调节非易失性存储器阵列的电导值；

具体的，通过在字线和位线上施加电压调整非易失性存储单元的权重，需要增加电导时，非易失性存储单元两侧电压等于 set 电压，需要减少电导时，非易失性存储单元两侧电压等于 reset 电压，直至调节为卷积核权重对应的电导值。

A4、将  $4 \times 4$  大小的电压矩阵输入到对应字线电极后，经过非易失性存储阵列在位线电极上得到  $2 \times 2$  大小的输出电流矩阵，即由 V11~V44 得到 I11~I33；

具体的，如图 6 中左下图所示，V41 施加在字线 1 上，按照①所示的电流通路经过 k21 后在位线 2 上得到电流，V42 施加在字线 2 上，按照②所示的电流通路经过 k22 后在位线 2 上得到电流，同理在图 7 中，V31 和 V32 分别经过 k11 和 k12 后在位线 2 上得到电流，四股电流在位线 2 上汇聚为 I31，即  $I_{31} = V_{31} \times k_{11} + V_{32} \times k_{12} + V_{41} \times k_{21} + V_{42} \times k_{22}$ 。同时，V42 还会经过 k21 在位线 3 上产生电流，即  $I_{32} = V_{32} \times k_{11} + V_{33} \times k_{12} + V_{42} \times k_{21} + V_{43} \times k_{22}$ ，V42 和 V32 参与了卷积运算中由于卷积核在 x 方向上滑动而产生的两次乘加运算，实现了输入数据的重复利用。同时在图 3 和图 7 中，V21~V24 和 V31~V34 会作用于字线电极相连的上下两层非易失性存储器阵列，参与了卷积运算中由于卷积核在 y 方向上滑动而产生的两次乘加运算，实现了输入数据的重复利用。 $4 \times 4$  的电压矩阵经过复制了多个  $2 \times 2$  卷积核的电导阵列后，得到  $3 \times 3$  的输出电流矩阵。

A5、输出电流矩阵经过输出信号读出装置线性转化为电压矩阵，将此电压传递给下一次卷积运算作为输入，或者储存此电压矩阵作为卷积运算的最终输出。

实施例 3、

为进一步说明本发明提出的基于三维非易失性存储器阵列的卷积运算方法以及其优势，下面以  $3 \times 3$  大小的卷积核为例，阐述本发明提出的卷积

运算方法的技术细节。

如图 8 所示为横向上具有八条字线的三维非易失性存储阵列的主视图，展示了  $3 \times 3$  大小的卷积核在多层非易失性存储单元层上的映射方案，输入数据为  $4 \times 4$  大小。将  $3 \times 3$  的卷积核切分为  $2 \times 3$  和  $1 \times 3$  的部分分别进行卷积运算，即图 8 中下方  $k_1$ 、 $k_2$  和  $k_3$ ， $k_1$ 、 $k_2$ 、 $k_3$  分别对应着  $3 \times 3$  大小卷积核的第一行、第二行和第三行； $V_1$ 、 $V_2$ 、 $V_3$ 、 $V_4$  对应  $4 \times 4$  输入的第一行、第二行、第三行和第四行。将  $V_1 \sim V_3$  以及  $k_1$ 、 $k_2$  映射到图 8 上方左边四条字线对应的阵列， $V_1 \sim V_3$  依次输入到第一层、第二层、第三层字线电极；第一层、第二层、第三层、第四层非易失存储单元层分别映射  $k_1$ 、 $k_2$ 、 $k_1$ 、 $k_2$ ，左边四条字线对应的阵列完成  $k_1$ 、 $k_2$  的卷积运算操作。将  $V_3$ 、 $V_4$  以及  $k_3$  映射到图 8 上方右边四条字线对应的阵列， $V_3$ 、 $V_4$  输入到第一层、第二层字线电极；第一层、第三层非易失存储单元层都映射  $k_3$ ，第二层、第四层非易失存储单元层的非易失性存储单元全部置为高阻态，左边四条字线对应的阵列完成  $k_3$  的卷积运算操作。每层的映射方案在图 9 中展示。

图 9 为横向上具有八条字线的三维非易失性存储阵列的俯视剖视图，展示了  $3 \times 3$  大小的卷积核在一层非易失性存储单元层上的映射方案。 $V_{11} \sim V_{14}$  为第一行的输入，同时以第一行卷积核  $k_{11} \sim k_{13}$  以及第三行卷积核  $k_{31} \sim k_{33}$  为例，展示在一层非易失性存储单元层上的映射方案。在左边 4 条字线上输入  $V_{11} \sim V_{14}$ ，在左边 4 条字线对应的非易失性存储单元层中，在斜对角线上映射  $k_{11} \sim k_{13}$ ，并平行复制一次；在右边 4 条字线上输入  $V_{11} \sim V_{14}$ ，在右边 4 条字线对应的非易失性存储单元层中，在斜对角线上映射  $k_{31} \sim k_{33}$ ，并平行复制一次。

相关技术方案同实施例 1 和实施例 2，这里不做赘述。

综上所述，本发明提出了一种基于三维非易失性存储器阵列的卷积神经网络运算加速器及运算方法，三维非易失性存储器阵列由一层字线电极、一层非易失性存储单元、一层位线电极、一层平面非易失性存储单元阵列依

次堆叠而形成的基本单元叠加构造而成，字线电极作为卷积运算的输入端，经过对应的非易失存储器阵列进行卷积运算，在位线上进行电流的叠加后得到卷积运算的输出。本发明提出的三维非易失性存储阵列有独特的拓扑结构，能够天然匹配卷积运算中卷积核在输入数据上滑动进行乘加运算的过程，一次性对二维输入数据进行卷积运算，得到二维输出数据。与现有技术相比，能够不破坏输入数据矩阵的位置信息，且能够得到矩阵形式的数据输出，符合软件中进行卷积运算的逻辑，同时并行度高，集成度高，在更小的占用面积下获得更多的运算次数。

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。

## 权 利 要 求

1、一种卷积运算加速器，其特征在于，包括：三维非易失性存储器阵列和控制模块；

所述三维非易失性存储器阵列包括： $P$  个字线电极层；任意相邻的两个字线电极层之间均放置有一个位线电极层；任意相邻的字线电极层和位线电极层之间均放置有一个非易失性存储单元阵列，且非易失性存储单元阵列与字线电极层和位线电极层均垂直相连；

所述字线电极层包括多个平行排布的字线电极；所述  $P$  个字线电极层中的字线电极共同构成一个字线电极阵列；

非易失性存储单元阵列中的每一列非易失性存储单元均连接在与非易失性存储单元阵列相连的字线电极层中的同一条字线上；非易失性存储单元阵列中每一条斜线上的非易失性存储单元均连接在与非易失性存储单元阵列相连的位线电极阵列中的同一条位线上；所述斜线为非易失性存储单元阵列中与非易失性存储单元阵列中的对应对角线平行的斜线；

记二维输入数据的尺寸为  $M \times N$ ；

当与所述二维输入数据进行卷积运算的卷积核尺寸为  $2k \times c$  时，所述控制模块用于将所述卷积核按行拆分为  $k$  个  $2 \times c$  大小的卷积核单元， $k$  为正整数；在字线电极阵列中选中  $k$  个大小为  $(M-2(k-1)) \times N$  的不同子阵列单元，并与所述  $k$  个  $2 \times c$  大小的卷积核单元按照卷积核的拆分顺序一一对应；将每个卷积核单元及其  $N-c$  次复制均存储至对应子阵列单元中所有相邻的两个字线电极层之间的两层非易失性存储单元阵列中；将所述二维输入数据上第  $2i-1$  行到第  $M-2(k-i)$  行的数据以电压的方式按照对应的坐标信息施加到第  $i$  个子阵列单元中的对应字线电极上， $i = 1, 2, \dots, k$ ；

当与所述二维输入数据进行卷积运算的卷积核尺寸为  $(2k+1) \times c$  时，所述控制模块用于将卷积核按行拆分为  $k$  个  $2 \times c$  大小的卷积核单元和一个

$1 \times c$  大小的卷积核单元；在字线电极阵列中选中  $k$  个大小为  $(M-2(k-1)-1) \times N$  的子阵列单元，以及一个大小为  $(M-2k) \times N$  的子阵列单元；所述  $k$  个大小为  $(M-2(k-1)-1) \times N$  的子阵列单元与所述  $k$  个  $2 \times c$  大小的卷积核单元按照卷积核的拆分顺序一一对应；将每个  $2 \times c$  大小的卷积核单元及其  $N-c$  次复制均存储至对应子阵列单元中所有相邻的两个字线电极层之间的两层非易失性存储单元阵列中；对所述大小为  $(M-2k) \times N$  的子阵列单元中的各字线电极层，在与其相连的其中一个非易失性存储单元阵列中分别存储所述  $1 \times c$  大小的卷积核单元及所述  $1 \times c$  大小的卷积核单元的  $N-c$  次复制，未存储卷积核单元的非易失性存储单元阵列中的非易失存储单元全部置为高阻态；将所述二维输入数据上第  $2i-1$  行到第  $M-2(k-i)-1$  行的数据以电压的方式按照对应的坐标信息施加到第  $i$  个  $(M-2(k-1)-1) \times N$  大小的子阵列单元中的对应字线电极上， $i = 1, 2, \dots, k$ ；将所述二维输入数据上第  $2k+1$  行到第  $M$  行的数据以电压的方式按照对应的坐标信息施加到所述大小为  $(M-2k) \times N$  的子阵列单元中的对应字线电极上；

所述三维非易失性存储器阵列用于基于非易失性存储单元阵列并行地实现卷积核单元与二维输入数据不同部分的点乘运算，经由对应的位线电极层并行输出卷积核单元与二维输入数据对应部分的点乘运算结果之和，从而实现所述卷积核与所述二维输入数据的卷积运算；

其中，卷积核单元的每一行卷积核数据在对应非易失性存储单元阵列的对应斜线上依次存储后，沿着卷积核滑动方向水平移动到相邻的  $N-c$  条斜线上再次对该行卷积核数据进行存储，实现每一层非易失性存储单元阵列存储卷积核单元中对应行的卷积核数据及该卷积核数据的  $N-c$  次复制。

2、根据权利要求 1 所述的卷积运算加速器，其特征在于，包括：当所述卷积核的尺寸为  $2k \times c$  时，所述三维非易失性存储器阵列用于基于所述大小为  $(M-2(k-1)) \times N$  的子阵列单元中所有相邻的两个字线电极层之间的两层非易失性存储单元阵列，并行地实现所述  $2 \times c$  大小的卷积核单元与二维输

入数据不同部分的点乘运算后,经由对应的位线电极层并行输出所述  $2 \times c$  大小的卷积核单元与所述二维输入数据对应部分的点乘运算结果之和;所述控制模块还用于对所选中的各子阵列单元中同一相对位置的位线输出分别进行汇总,得到卷积核与二维输入数据的卷积运算结果;

当所述卷积核的尺寸为  $(2k+1) \times c$  时,所述三维非易失性存储器阵列用于基于所述大小为  $(M-2(k-1)-1) \times N$  的子阵列单元中所有相邻的两个字线电极层之间的两层非易失性存储单元阵列,并行地实现  $2 \times c$  大小的卷积核单元与所述二维输入数据不同部分的点乘运算后,经由对应的位线电极层并行输出所述  $2 \times c$  大小的卷积核单元与所述二维输入数据对应部分的点乘运算结果之和;同时基于所述大小为  $(M-2k) \times N$  的子阵列单元中的非易失性存储单元阵列,并行地实现所述  $1 \times c$  大小的卷积核单元与所述二维输入数据不同部分的点乘运算后,经由对应的位线电极层并行输出  $1 \times c$  大小的卷积核单元与二维输入数据对应部分的点乘运算结果之和;所述控制模块还用于将所选中的各子阵列单元中的各位线输出对应进行汇总,得到卷积核与二维输入数据的卷积运算结果。

3、根据权利要求 1 所述的卷积运算加速器,其特征在于,所述三维非易失性存储器阵列为多层堆叠结构,由字线电极层、所述非易失性存储单元阵列和位线电极层依次堆叠形成;各字线电极层之间、字线电极层与位线电极层之间均互相平行。

4、根据权利要求 1 所述的卷积运算加速器,其特征在于,所述控制模块包括:输入端信号发生装置、输出端信号发生装置和输出信号读出装置;

所述输入端信号发生装置和输出端信号发生装置相互配合,在子阵列单元所对应的字线电极和位线电极上施加电压,选择对应的非易失存储单元,并改变所选中的各非易失存储单元的电导值,使其存储对应的卷积核数据;

所述输入端信号发生装置还用于将二维输入数据以电压的方式,按照

二维输入数据的坐标信息施加到子阵列单元中的对应字线电极上；

所述输出信号读出装置用于检测位线电极上的输出电流，并转化为电压信号，作为卷积运算的结果。

5、根据权利要求4所述的卷积运算加速器，其特征在于，所述输入信号发生装置的读电压小于或等于非易失存储单元上的当前 set 电压值，以保证在进行卷积运算时非易失存储单元的电导值不被改变。

6、根据权利要求1-5任意一项所述的卷积运算加速器，其特征在于，所述卷积运算加速器用于实现卷积神经网络中的卷积运算。

7、根据权利要求6所述的卷积运算加速器，其特征在于，在卷积神经网络的训练过程中，所述控制模块还用于通过在对应的字线电极和位线电极上施加写电压来调整各非易失存储单元的电导值，以对卷积神经网络的参数进行更新，并将卷积运算加速器上一轮的卷积运算结果重新作为卷积运算加速器下一轮卷积运算的输入。

8、根据权利要求7所述的卷积运算加速器，其特征在于，当需要增加非易失存储单元的电导值时，在对应的字线电极和位线电极上施加的写电压差值的绝对值为对应的 set 电压值；

当需要减小非易失存储单元的电导值时，在对应的字线电极和位线电极上施加的写电压差值的绝对值为对应的 reset 电压值。

9、一种基于权利要求1-8任意一项所述卷积运算加速器的卷积运算方法，其特征在于，包括以下步骤：

S1、当与所述二维输入数据进行卷积运算的卷积核尺寸为  $2k \times c$  时，将所述卷积核按行拆分为  $k$  个  $2 \times c$  大小的卷积核单元， $k$  为正整数；在字线电极阵列中选中  $k$  个大小为  $(M-2(k-1)) \times N$  的不同子阵列单元，并与所述  $k$  个  $2 \times c$  大小的卷积核单元按照卷积核的拆分顺序一一对应；将每个卷积核单元及其  $N-c$  次复制均存储至对应子阵列单元中所有相邻的两个字线电极层之间的两层非易失性存储单元阵列中；将所述二维输入数据上第  $2i-1$  行到第

$M-2(k-i)$ 行的数据以电压的方式按照对应的坐标信息施加到第  $i$  个子阵列单元中的对应字线电极上,  $i = 1, 2, \dots, k$ ;

当与所述二维输入数据进行卷积运算的卷积核尺寸为  $(2k+1) \times c$  时, 将卷积核按行拆分为  $k$  个  $2 \times c$  大小的卷积核单元和一个  $1 \times c$  大小的卷积核单元; 在字线电极阵列中选中  $k$  个大小为  $(M-2(k-1)-1) \times N$  的子阵列单元, 以及一个大小为  $(M-2k) \times N$  的子阵列单元; 所述  $k$  个大小为  $(M-2(k-1)-1) \times N$  的子阵列单元与所述  $k$  个  $2 \times c$  大小的卷积核单元按照卷积核的拆分顺序一一对应; 将每个  $2 \times c$  大小的卷积核单元及其  $N-c$  次复制均存储至对应子阵列单元中所有相邻的两个字线电极层之间的两层非易失性存储单元阵列中; 对所述大小为  $(M-2k) \times N$  的子阵列单元中的各字线电极层, 在与其相连的其中一个非易失性存储单元阵列中分别存储所述  $1 \times c$  大小的卷积核单元及所述  $1 \times c$  大小的卷积核单元的  $N-c$  次复制, 未存储卷积核单元的非易失性存储单元阵列中的非易失存储单元全部置为高阻态; 将所述二维输入数据上第  $2i-1$  行到第  $M-2(k-i)-1$  行的数据以电压的方式按照对应的坐标信息施加到第  $i$  个  $(M-2(k-1)-1) \times N$  大小的子阵列单元中的对应字线电极上,  $i = 1, 2, \dots, k$ ; 将所述二维输入数据上第  $2k+1$  行到第  $M$  行的数据以电压的方式按照对应的坐标信息施加到所述大小为  $(M-2k) \times N$  的子阵列单元中的对应字线电极上;

S2、基于非易失性存储单元阵列并行地实现卷积核单元与二维输入数据不同部分的点乘运算, 经由对应的位线电极层并行输出卷积核单元与二维输入数据对应部分的点乘运算结果之和, 从而实现所述卷积核与所述二维输入数据的卷积运算;

其中, 所述二维输入数据的尺寸为  $M \times N$ ; 卷积核单元的每一行卷积核数据在对应非易失性存储单元阵列的对应斜线上依次存储后, 沿着卷积核滑动方向水平移动到相邻的  $N-c$  条斜线上再次对该行卷积核数据进行存储,

实现每一层非易失性存储单元阵列存储卷积核单元中对应行的卷积核数据及该卷积核数据的  $N-c$  次复制。

10、根据权利要求 9 所述的卷积运算方法，其特征在于，当所述卷积核的尺寸为  $2k \times c$  时，所述步骤 S2 包括：基于所述大小为  $(M-2(k-1)) \times N$  的子阵列单元中所有相邻的两个字线电极层之间的两层非易失性存储单元阵列，并行地实现所述  $2 \times c$  大小的卷积核单元与二维输入数据不同部分的点乘运算后，经由对应的位线电极层并行输出所述  $2 \times c$  大小的卷积核单元与所述二维输入数据对应部分的点乘运算结果之和；对所选中的各子阵列单元中同一相对位置的位线输出分别进行汇总，得到所述卷积核与所述二维输入数据的卷积运算结果；

当所述卷积核的尺寸为  $(2k+1) \times c$  时，所述步骤 S2 包括：基于所述大小为  $(M-2(k-1)-1) \times N$  的子阵列单元中所有相邻的两个字线电极层之间的两层非易失性存储单元阵列，并行地实现  $2 \times c$  大小的卷积核单元与所述二维输入数据不同部分的点乘运算后，经由对应的位线电极层并行输出所述  $2 \times c$  大小的卷积核单元与所述二维输入数据对应部分的点乘运算结果之和；同时基于所述大小为  $(M-2k) \times N$  的子阵列单元中的非易失性存储单元阵列，并行地实现所述  $1 \times c$  大小的卷积核单元与所述二维输入数据不同部分的点乘运算后，经由对应的位线电极层并行输出  $1 \times c$  大小的卷积核单元与二维输入数据对应部分的点乘运算结果之和；将所选中的各子阵列单元中的各位线输出对应进行汇总，得到所述卷积核与所述二维输入数据的卷积运算结果。

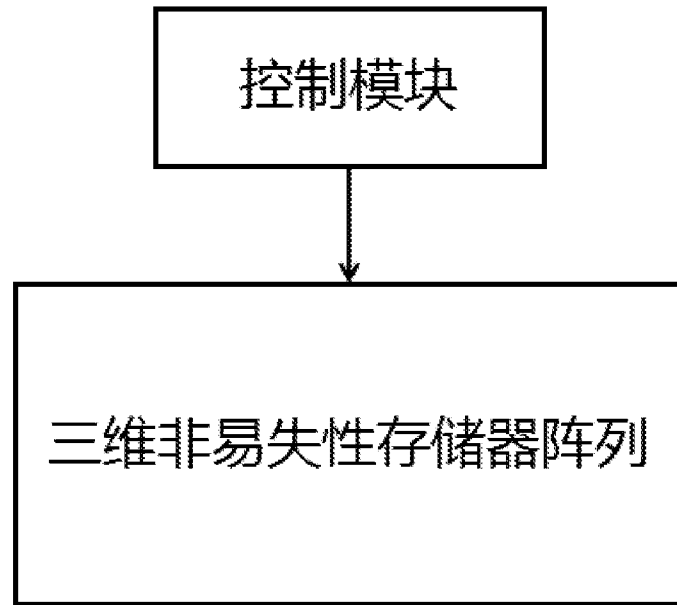


图 1

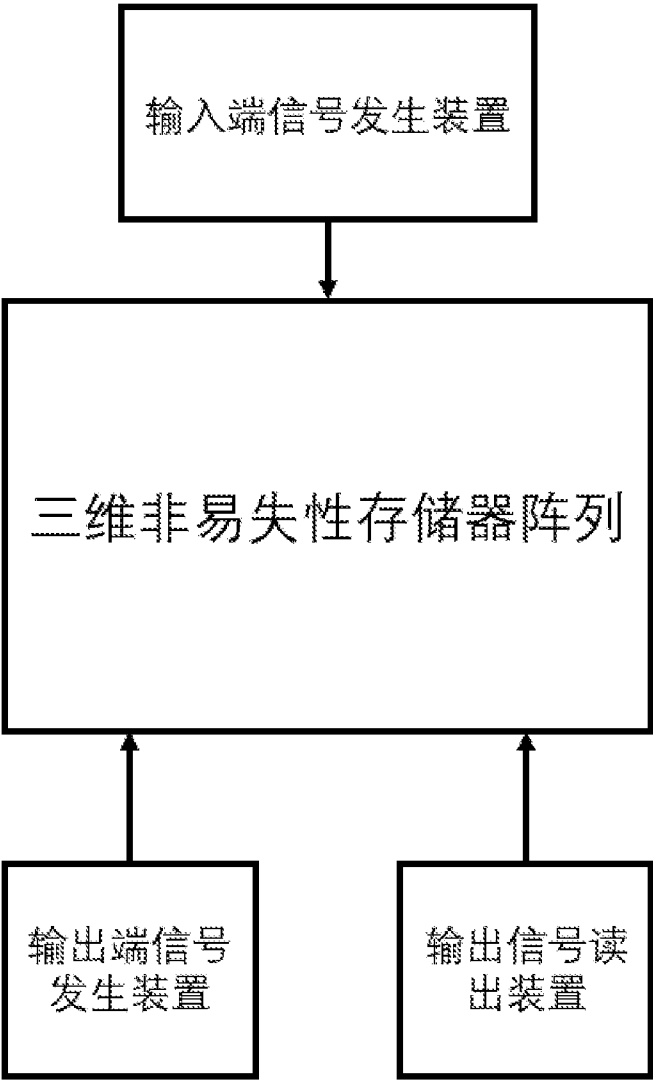


图 2

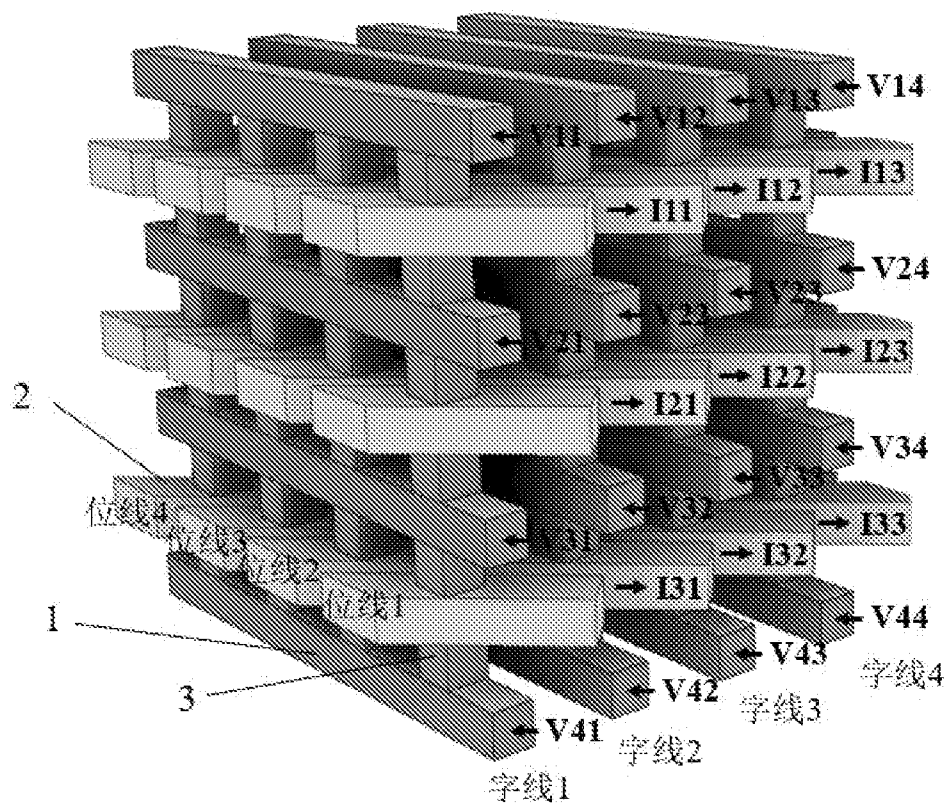


图 3

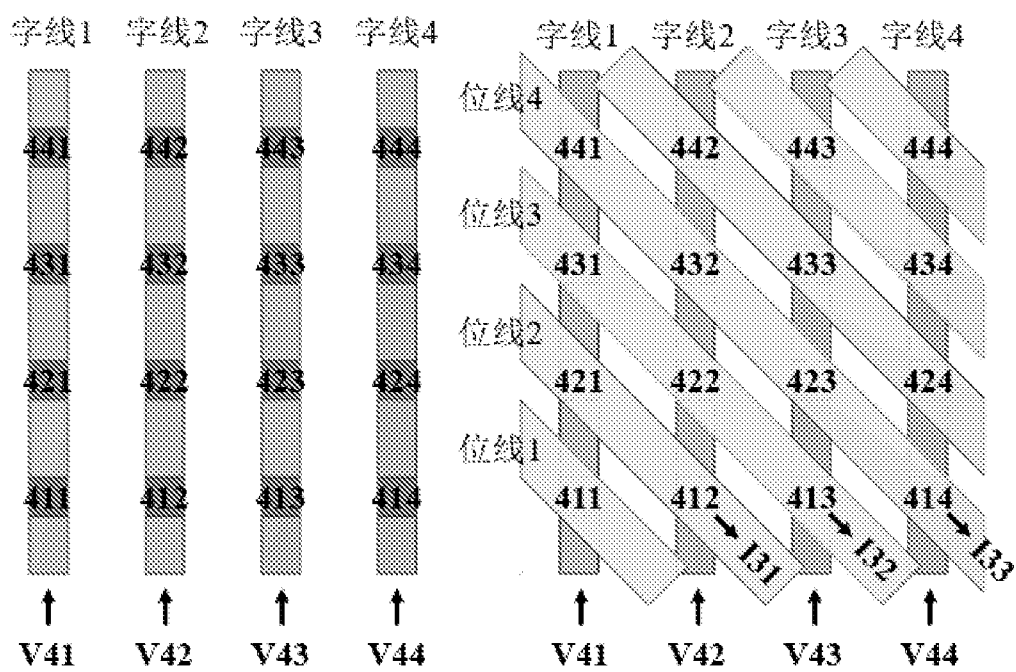


图 4

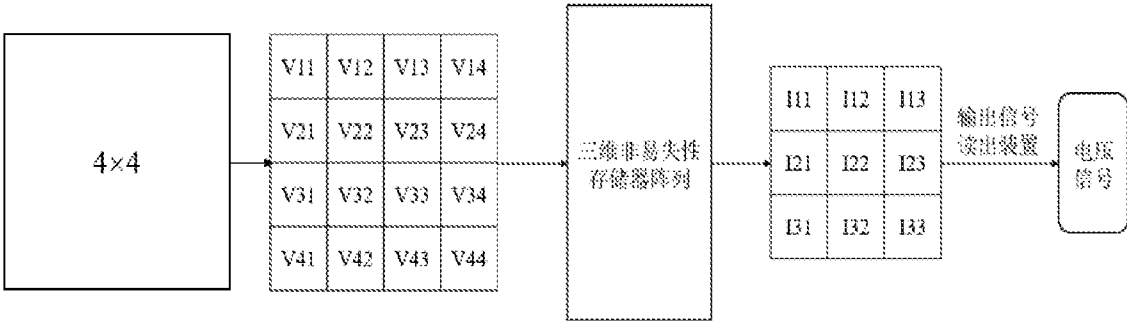


图 5

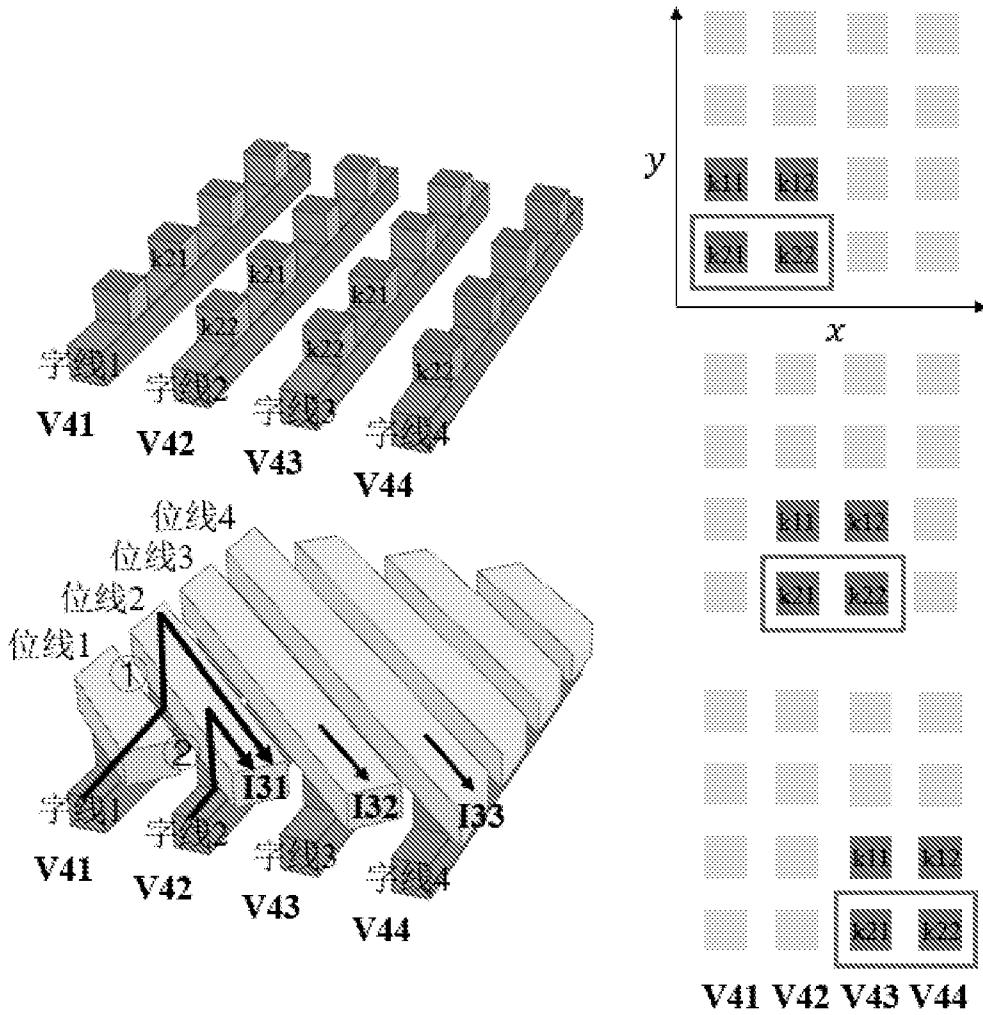


图 6

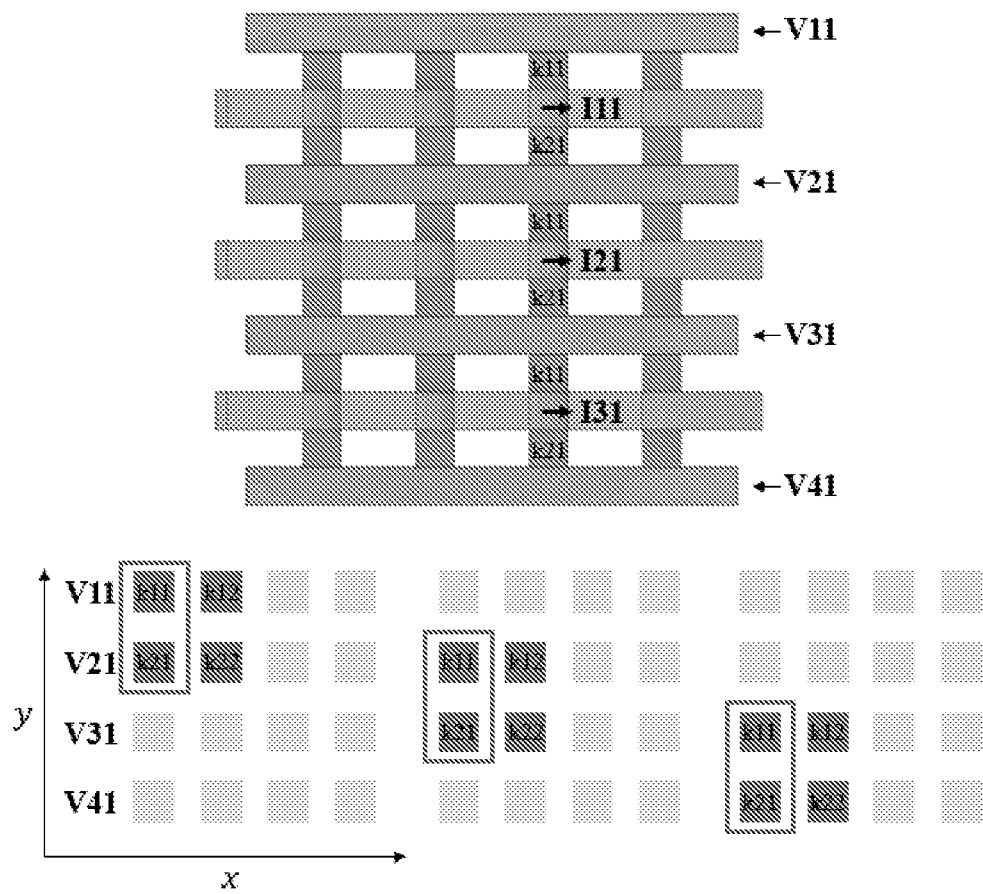


图 7

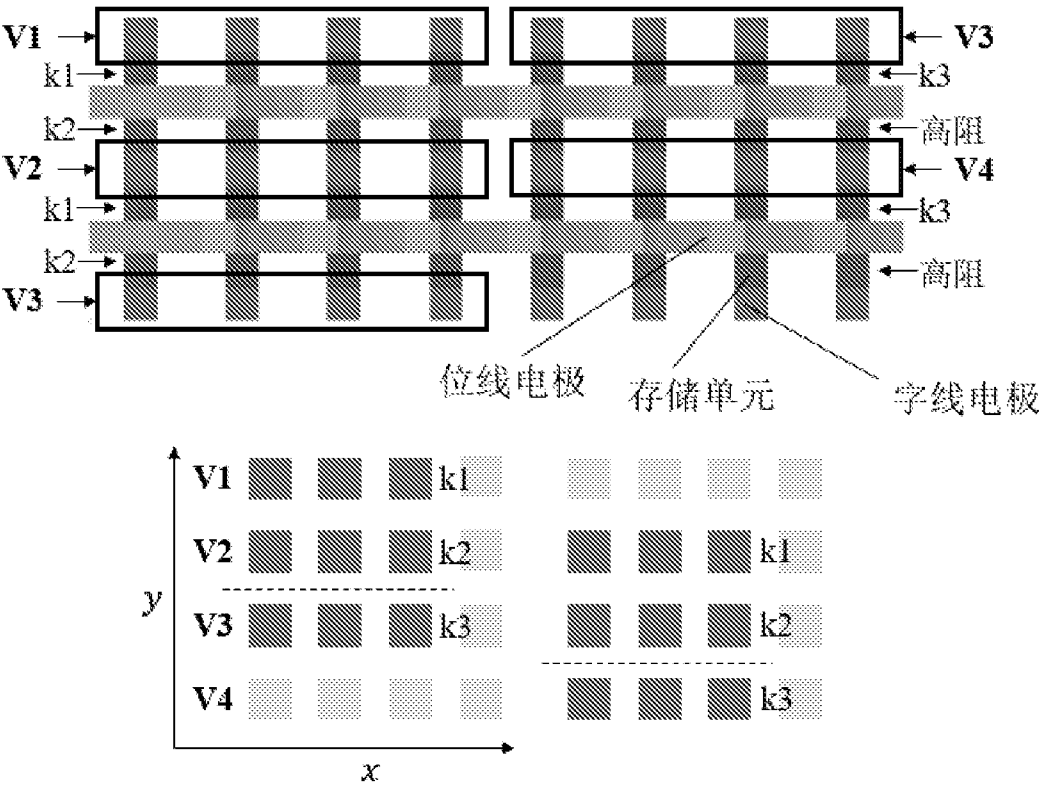


图 8

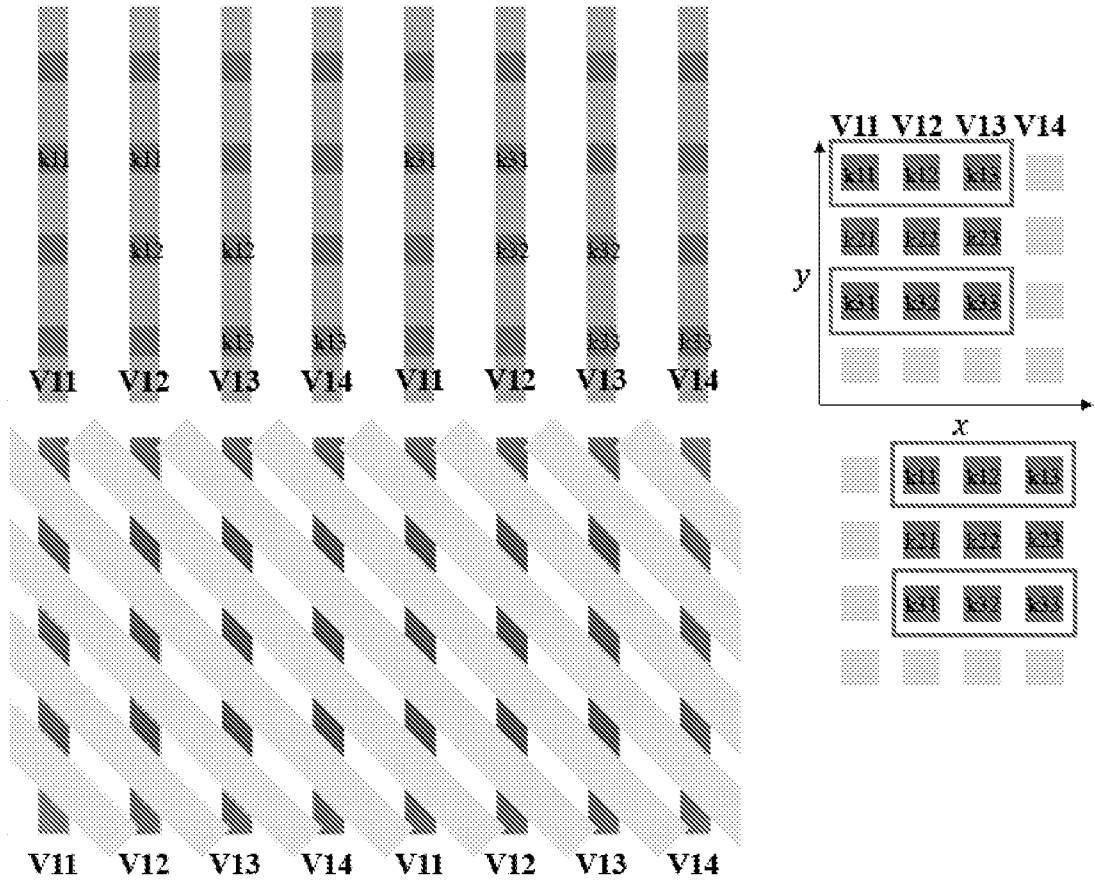


图 9

## INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2022/087794

**A. CLASSIFICATION OF SUBJECT MATTER**

G11C 5/02(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)

G11C

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, BAIDU, CNPAT, CNKI, IEEE: 卷积, 加快, 加速, 并行, 三维, 存储器阵列, 滑动, 拆分, convolutional, accelerated, parallel, three-Dimensional, memory array, slide, splitting

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

| Category* | Citation of document, with indication, where appropriate, of the relevant passages                                                    | Relevant to claim No. |
|-----------|---------------------------------------------------------------------------------------------------------------------------------------|-----------------------|
| A         | CN 110826709 A (HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY) 21 February 2020 (2020-02-21)<br>description, paragraphs [0033]-[0055] | 1-10                  |
| A         | CN 106530210 A (PEKING UNIVERSITY) 22 March 2017 (2017-03-22)<br>entire document                                                      | 1-10                  |
| A         | CN 107368889 A (PEKING UNIVERSITY) 21 November 2017 (2017-11-21)<br>entire document                                                   | 1-10                  |
| A         | CN 106847335 A (PEKING UNIVERSITY) 13 June 2017 (2017-06-13)<br>entire document                                                       | 1-10                  |
| A         | US 2019108437 A1 (HSU FU-CHANG et al.) 11 April 2019 (2019-04-11)<br>entire document                                                  | 1-10                  |



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 2022

Date of mailing of the international search report

26 October 2022

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/CN2022/087794**

| Patent document<br>cited in search report |            |    | Publication date<br>(day/month/year) | Patent family member(s) |            |    | Publication date<br>(day/month/year) |
|-------------------------------------------|------------|----|--------------------------------------|-------------------------|------------|----|--------------------------------------|
| CN                                        | 110826709  | A  | 21 February 2020                     | WO                      | 2021073104 | A1 | 22 April 2021                        |
| CN                                        | 106530210  | A  | 22 March 2017                        | None                    |            |    |                                      |
| CN                                        | 107368889  | A  | 21 November 2017                     | None                    |            |    |                                      |
| CN                                        | 106847335  | A  | 13 June 2017                         | None                    |            |    |                                      |
| US                                        | 2019108437 | A1 | 11 April 2019                        | CN                      | 109754065  | A  | 14 May 2019                          |

## 国际检索报告

国际申请号

PCT/CN2022/087794

## A. 主题的分类

G11C 5/02 (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)

G11C

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, BAIDU, CNPAT, CNKI, IEEE: 卷积, 加快, 加速, 并行, 三维, 存储器阵列, 滑动, 拆分, convolutional, accelerated, parallel, three-Dimensional, memory array, slide, splitting

## 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 |
|------|---------------------------------------------------------------------------|---------|
| A    | CN 110826709 A (华中科技大学) 2020年2月21日 (2020 - 02 - 21)<br>说明书第[0033]-[0055]段 | 1-10    |
| A    | CN 106530210 A (北京大学) 2017年3月22日 (2017 - 03 - 22)<br>全文                   | 1-10    |
| A    | CN 107368889 A (北京大学) 2017年11月21日 (2017 - 11 - 21)<br>全文                  | 1-10    |
| A    | CN 106847335 A (北京大学) 2017年6月13日 (2017 - 06 - 13)<br>全文                   | 1-10    |
| A    | US 2019108437 A1 (HSU, Fu-Chang 等) 2019年4月11日 (2019 - 04 - 11)<br>全文      | 1-10    |

☐ 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

2022年10月18日

国际检索报告邮寄日期

2022年10月26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

吴少鸿

电话号码 86-(10)-53961533

国际检索报告  
关于同族专利的信息

国际申请号

PCT/CN2022/087794

| 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          | 110826709  | A  | 2020年2月21日     | WO   | 2021073104 | A1 | 2021年4月22日     |
| CN          | 106530210  | A  | 2017年3月22日     | 无    |            |    |                |
| CN          | 107368889  | A  | 2017年11月21日    | 无    |            |    |                |
| CN          | 106847335  | A  | 2017年6月13日     | 无    |            |    |                |
| US          | 2019108437 | A1 | 2019年4月11日     | CN   | 109754065  | A  | 2019年5月14日     |