(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 110008440 B
(45)授权公告日 2021.07.27

(21)申请号 201910299610.9

(22)申请日 2019.04.15

(65)同一申请的已公布的文献号
申请公布号 CN 110008440 A

(43)申请公布日 2019.07.12

(73)专利权人 恒烁半导体(合肥)股份有限公司
地址 230000 安徽省合肥市庐阳区天水路与太和路交口西北庐阳中科大校友企业创新园11号楼

(72)发明人 任军 徐伟民 蒋明峰 李政达 吕向东 徐培

(74)专利代理机构 北京中政联科专利代理事务所(普通合伙) 11489
代理人 刘艳

(51)Int.Cl.
G06F 17/16(2006.01)
G06N 3/063(2006.01)

审查员 李兵兵

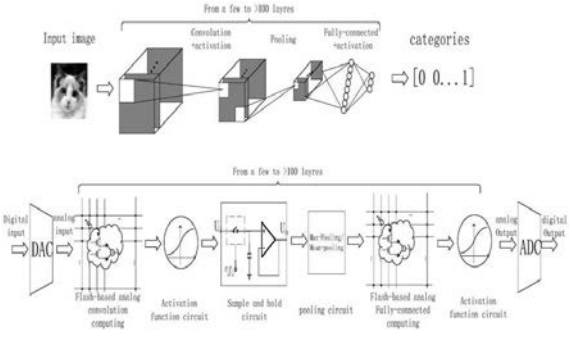
权利要求书3页 说明书13页 附图17页

(54)发明名称

一种基于模拟矩阵运算单元的卷积运算及其应用

(57)摘要

本发明涉及电路设计技术领域，公开了一种基于模拟矩阵运算单元的卷积运算及其应用，所述卷积运算将卷积核经过转换拼接成P个长度为Q的横向量，并映射到模拟矩阵运算单元中，输入特征图按照滑动窗口进行切分转换成N个长度为Q的横向量，在脉冲波下，依次将N个长度为Q的横向量映射到向量中，同时输出N个长度为P的运算结果经采样保持器得到完整的卷积结果输出。本发明有效提高了卷积计算的速率，减少了功耗和电路面积，具有高度计算并行性，大大提高了计算的密度和效率，具有较高的实用价值和广泛的应用前景。



CN 110008440 B

1.一种基于模拟矩阵运算单元的卷积运算方法，其特征在于，所述模拟矩阵运算单元能够实现向量$(a_1 \cdots a_Q)$与矩阵$\begin{pmatrix} b_{11} & \cdots & b_{1Q} \\ \vdots & \ddots & \vdots \\ b_{P1} & \cdots & b_{PQ} \end{pmatrix}^T$的乘法输出；

所述卷积运算包括以下步骤：

将卷积核经过转换拼接成P个长度为Q的横向量，并映射到矩阵$\begin{pmatrix} b_{11} & \cdots & b_{1Q} \\ \vdots & \ddots & \vdots \\ b_{P1} & \cdots & b_{PQ} \end{pmatrix}$中；

将输入特征图按照滑动窗口进行切分转换成N个长度为Q的横向量；

在N个脉冲波下，依次将N个长度为Q的横向量映射到向量$(a_1 \cdots a_Q)$中，矩阵运算单元的模拟输出端口按照脉冲波序列，输出N个长度为P的运算结果；

通过采样保持器将所有输出的运算结果采样保持至一个时间点，得到完整的卷积结果；

所述模拟矩阵运算单元由P个模拟乘加单元构成，所述模拟乘加单元能够实现一个行向量$[V_1, V_2, \ldots, V_n]$与列向量$[W_1, W_2, \ldots, W_n]^T$的乘法输出；

所述模拟乘加单元由Q个模拟乘法电路横构成；

所述模拟乘法电路包括一对浮栅场效应管M1、M2和一个差分电流检测电路；所述M_1和M_2共栅极并接入电压源，共漏极或共源极并接入模拟电压输入，所述差分电流检测电路包括两个电流输入端和一个输出端，所述两个电流输入端对应接入M_1和M_2各自源极或漏极，所述差分电流检测电路能保持两个电流输入端电压不变且输出端结果为两个电流输入端输入电流差值的函数；所述M_1和M_2的栅源电压大于两者阈值电压的最大值，漏源电压小于栅源电压分别与两者阈值电压差值的最小值；

所述模拟乘加单元中的Q个模拟乘法电路共栅极且共用一个差分电流检测电路，所述每个模拟乘法电路中该对浮栅场效应管的共漏极或共源极分别接入对应的输入电压信号；所述差分电流检测电路的输出端结果为每个模拟乘法电路输出的基于该对浮栅场效应管输出电流差值的函数之和；

所述模拟乘法电路执行乘法运算的步骤包括：

对两个浮栅场效应管M_1、M_2中的浮置栅极执行擦除和写操作，将乘数以所述一对浮栅场效应管的阈值电压差值的形式进行存储；

对两个浮栅场效应管源极、漏极和选栅极执行电压施加，以所述模拟电压输入与所述两个电流输入端电压差值的形式输入被乘数；

通过差分电流检测电路输出基于所述一对浮栅场效应管输出电流差值的函数作为所述被乘数与乘数的乘积结果。

2.根据权利要求1所述的一种基于模拟矩阵运算单元的卷积运算方法，其特征在于，所述模拟矩阵运算单元为一个P行Q列的模块电路，所述模块电路中的每列浮栅场效应管的共漏极或共源极接入同一的输入电压信号；

通过控制P个模拟乘加单元的栅极电压，确定参加计算的乘加单元的个数，实现向量

$[a_1 \cdots a_Q]$与矩阵$\begin{pmatrix} b_{11} & \cdots & b_{1Q} \\ \vdots & \ddots & \vdots \\ b_{J1} & \cdots & b_{JQ} \end{pmatrix}^T$的乘法输出，其中J≤P。

3.根据权利要求2所述的一种基于模拟矩阵运算单元的卷积运算方法，其特征在于，所述模拟矩阵运算单元在确定参加计算的乘加单元的个数后执行以下步骤：

对每对浮栅场效应管中的浮置栅极执行擦除和写操作，按照一对浮栅场效应管的阈值电压差值对应一个乘数进行存储，完成大小为J×Q的乘数矩阵的存储；

对每对浮栅场效应管源极、漏极和控制栅极执行电压施加，以每个乘法电路中模拟电压输入与所述两个电流输入端电压差值对应一个被乘数，完成元素个数为Q的横向被乘数向量的输入；

通过每行中设置差分电流检测电路输出得到被乘数向量与乘数矩阵的乘积结果。

4.根据权利要求1所述的一种基于模拟矩阵运算单元的卷积运算方法，其特征在于，所述差分电流检测电路包括两个电流-电压转换器和一个电压差分放大器；所述电流-电压转换器由一个运算放大器和一个阻性负载构成，运算放大器的负极输入端为所述电流-电压转换器的电流输入端口，运算放大器的输出端为所述电流-电压转换器的电压输出端口，运算放大器的正极输入端外加直流电平V_b，所述阻性负载连接在运算放大器的负极输入端与运算放大器的输出端之间。

5.根据权利要求1所述的一种基于模拟矩阵运算单元的卷积运算方法，其特征在于，所述电压差分放大器由一个运算放大器、两个阻值相同的电阻R_1、R_3和两个阻值相同的电阻R_2、R_4构成，所述电压差分放大器的两个电压输入端口分别通过电阻R_1、R_3连接到运算放大器的正、负极输入端，运算放大器的正极输入端通过电阻R_2连接到直流电平V_b，运算放大器的负极输入端通过电阻R_4连接到运算放大器的输出端，运算放大器的输出端为所述电压差分放大器的电压输出端口。

6.根据权利要求4所述的一种基于模拟矩阵运算单元的卷积运算方法，其特征在于，所述阻性负载为浮栅场效应管或MOS管，所述浮栅场效应管或MOS管的栅极外加栅极电压，漏极或源极连接到所述运算放大器的输出端，源极或漏极连接到所述运算放大器的负极输入端；

所述浮栅场效应管或MOS管的漏源电压V_{DS}、栅源电压V_{GS}和阈值电压V_{TH}满足：$V_{DS} \ll 2(V_{GS}-V_{TH})$。

7.一种采用如权利要求1-6任意一项所述的基于模拟矩阵运算单元的卷积运算方法实现卷积神经网络的方法，其特征在于，包括：

DAC电路、卷积层电路、激活函数电路、采样保持电路、池化层电路、全连接层电路和ADC电路，各个电路间的级联构成卷积神经网络；

所述的DAC电路用于将输入层的数字输入转化为模拟输入；

所述卷积层电路与全连接层电路均采用所述基于模拟矩阵运算单元的卷积运算实现功能并构建电路；

所述的激活函数电路将卷积层的输出电流转换为电压，同时完成激活函数作用；

所述的采样保持电路用于将卷积层的序列计算结果采样并保持在同一时间点；

所述的池化层电路用于池化层的实现，包括平均池化电路和最大池化电路；

所述的ADC电路用于将输出层的模拟输出转换为数字输出；

具体步骤包括：

输入特征图经过DAC电路转换为模拟信号输入；

经过卷积层电路得到第一次卷积计算结果，然后经过激活函数电路，将计算结果从电流转变为电压，并且进行非线性化；

经过采样保持电路，得到一个完整的中间特征图，再经过池化电路，将模拟信号降维；

继续经过全连接层电路进行全连接运算，最后经过激活函数电路得到模拟输出特征图，并经ADC电路转换为数字输出。

8.根据权利要求7所述的一种实现卷积神经网络的方法，其特征在于，所述全连接层电路为一个P行Q列的模拟矩阵运算单元，具体实现方法为：

将输入层$[a_1 \cdots a_Q]$按照每个乘法电路中模拟电压输入与两个电流输入端电压差值对应一个a_i其中$i \in (1 \cdots Q)$，从模拟矩阵运算单元模拟电压输入端输入；

将全连接层的$P \times Q$个权重拆分为P个横向量，即$[w_{11}, w_{21} \cdots w_{Q1}]$，$[w_{12}, w_{22}, \cdots w_{Q2}]$，……$[w_{1P}, w_{2P}, \cdots w_{QP}]$，按照一对浮栅场效应管的阈值电压差值对应一个乘数进行存储，将其映射到模拟矩阵运算单元的P行中，完成大小为$P \times Q$的权重矩阵的存储；

经过一个计算周期，通过采样保持器将所有输出的运算结果采样保持至一个时间点，得到完整的卷积结果，输出层向量$[b_1, b_2, \cdots b_P]$。

一种基于模拟矩阵运算单元的卷积运算及其应用

技术领域

[0001] 本发明涉及电路设计技术领域，具体涉及一种基于模拟矩阵运算单元的卷积运算及 其应用。

背景技术

[0002] 卷积神经网络在图像识别，目标检测和许多机器学习应用领域显示出巨大的优势。卷积神经网络主要由卷积层，池化层，全连接层级联组成，主要有输入层像素块和卷积核之间的卷积操作、为引入非线性而进行的激活操作、为减少特征值而对特征图进行的 下采样操作(即池化)以及卷积之后的全连接操作，其中，绝大部分计算量都在卷积层 和全连接层。

[0003] 大型的卷积神经网络，具有庞大的参数集和计算量。为完成庞大的计算量，一般芯片的设计思路是基于冯诺依曼架构上大量增加并行的运算单元，从早期的GPU，再到现 在的FPGA，ASIC，NPU，TPU，都是由控制单元，存储单元，计算单元构成的。运算过 程中，首先要将权重和输入特征存至片外的存储器，然后将需要运算的数据通过片内的 一二级缓存，再到寄存器，最后送入ALU单元进行运算。这种架构存在两大问题：运算 过程中，片内片外的数据往返传输消耗了大量的运算时间和功耗，计算单元和存储器之 间的数据搬运消耗了大量的资源；为满足算力需要不断增加的并行计算单元的数量与存 储单元带宽之间存在矛盾，成为AI芯片算力提升的瓶颈。

发明内容

[0004] 针对现有技术的不足，本发明提供一种模拟乘法电路、模拟乘法方法及其应用，用以解决背景技术中提出的问题。

[0005] 本发明解决技术问题采用如下技术方案：

[0006] 一种基于模拟矩阵运算单元的卷积运算，所述模拟矩阵运算单元能够实现向量$(a_1 \cdots a_Q)$与矩阵$\begin{pmatrix} b_{11} & \cdots & b_{1Q} \\ \vdots & \ddots & \vdots \\ b_{P1} & \cdots & b_{PQ} \end{pmatrix}^T$的乘法输出；

[0007] 所述卷积运算包括以下步骤：

[0008] 将卷积核经过转换拼接成P个长度为Q的横向量，并映射到矩阵$\begin{pmatrix} b_{11} & \cdots & b_{1Q} \\ \vdots & \ddots & \vdots \\ b_{P1} & \cdots & b_{PQ} \end{pmatrix}$中；

[0009] 将输入特征图按照滑动窗口进行切分转换成N个长度为Q的横向量；

[0010] 在N个脉冲波下，依次将N个长度为Q的横向量映射到向量$(a_1 \cdots a_Q)$中，矩阵运 算单元的模拟输出端口按照脉冲波序列，输出N个长度为P的运算结果；

[0011] 通过采样保持器将所有输出的运算结果采样保持至一个时间点，得到完整的卷积

结 果。

[0012] 优选地，所述模拟矩阵运算单元由P个模拟乘加单元构成，所述模拟乘加单元能够实现一个行向量$[V_1,V_2,\cdots,V_n]$与列向量$[W_1,W_2,...,W_n]^T$的乘法输出。

[0013] 优选地，所述模拟乘加单元由Q个模拟乘法电路横构成；

[0014] 所述模拟乘法电路包括一对浮栅场效应管M1、M2和一个差分电流检测电路；所述M_1和M_2共栅极并接入电压源，共漏极或共源极并接入模拟电压输入，所述差分电流检 测电路包括两个电流输入端和一个输出端，所述两个电流输入端对应接入M_1和M_2各自 源极或漏极，所述差分电流检测电路能保持两个电流输入端电压不变且输出端结果为两 个电流输入端输入电流差值的函数；所述M_1和M_2的栅源电压大于两者阈值电压的最大 值，漏源电压小于栅源电压分别与两者阈值电压差值的最小值；

[0015] 所述模拟乘加单元中的Q个模拟乘法电路共栅极且共同一个差分电流检测电路，所 述每个模拟乘法电路中该对浮栅场效应管的共漏极或共源极分别接入对应的输入电压信号；所述差分电流检测电路的输出端结果为每个模拟乘法电路输出的基于该对浮栅场效应管输出电流差值的函数之和。

[0016] 优选地，所述模拟乘法电路执行乘法运算的步骤包括：

[0017] 对两个浮栅场效应管M_1、M_2中的浮置栅极执行擦除和写操作，将乘数以所述一对浮栅场效应管的阈值电压差值的形式进行存储；

[0018] 对两个浮栅场效应管源极、漏极和选栅极执行电压施加，以所述模拟电压输入与所 述两个电流输入端电压差值的形式输入被乘数；

[0019] 通过差分电流检测电路输出基于所述一对浮栅场效应管输出电流差值的函数作为 所述被乘数与乘数的乘积结果。

[0020] 优选地，所述模拟矩阵运算单元为一个P行Q列的模块电路，所述模块电路中的每列浮栅场效应管的共漏极或共源极接入同一的输入电压信号；

[0021] 通过控制P个模拟乘加单元的栅极电压，确定参加计算的乘加单元的个数，实现向量$[a_1\cdots a_Q]$与矩阵$\begin{pmatrix} b_{11} & \cdots & b_{1Q} \\ \vdots & \ddots & \vdots \\ b_{J1} & \cdots & b_{JQ} \end{pmatrix}^T$的乘法输出，其中J≤P。

[0022] 优选地，所述模拟矩阵运算单元在确定参加计算的乘加单元的个数后执行以下步骤：

[0023] 对每对浮栅场效应管中的浮置栅极执行擦除和写操作，按照一对浮栅场效应管的阈 值电压差值对应一个乘数进行存储，完成大小为J×Q的乘数矩阵的存储；

[0024] 对每对浮栅场效应管源极、漏极和控制栅极执行电压施加，以每个乘法电路中模拟 电压输入与所述两个电流输入端电压差值对应一个被乘数，完成元素个数为Q的横向被乘数向量的输入；

[0025] 通过每行中设置差分电流检测电路输出得到被乘数向量与乘数矩阵的乘积结果。

[0026] 优选地，所述差分电流检测电路包括两个电流-电压转换器和一个电压差分放大器；所述电流-电压转换器由一个运算放大器和一个阻性负载构成，运算放大器的负极输入 端为所述电流-电压转换器的电流输入端口，运算放大器的输出端为所述电流-电压转

换 器的电压输出端口，运算放大器的正极输入端外加直流电平V_b，所述阻性负载连接在运算放大器的负极输入端与运算放大器的输出端之间。

[0027] 优选地，所述电压差分放大器由一个运算放大器、两个阻值相同的电阻R_1、R_3和两个阻值相同的电阻R_2、R_4构成，所述电压差分放大器的两个电压输入端口分别通过电阻 R_1、R_3连接到运算放大器的正、负极输入端，运算放大器的正极输入端通过电阻R_2连接 到直流电平V_b，运算放大器的负极输入端通过电阻R_4连接到运算放大器的输出端，运 算放大器的输出端为所述电压差分放大器的电压输出端口。

[0028] 优选地，所述阻性负载为浮栅场效应管或MOS管，所述浮栅场效应管或MOS管的 栅极外加栅极电压，漏极或源极连接到所述运算放大器的输出端，源极或漏极连接到所 述运算放大器的负极输入端；

[0029] 所述浮栅场效应管或MOS管的漏源电压V_{DS}、栅源电压V_{GS}和阈值电压V_{TH}满足：$V_{DS}<<2(V_{GS}-V_{TH})$。

[0030] 本发明还提供一种采用所述的基于模拟矩阵运算单元的卷积运算实现卷积神经网 络的方法，包括：

[0031] DAC电路、卷积层电路、激活函数电路、采样保持电路、池化层电路、全连接层电 路和ADC电路，各个电路间的级联构成卷积神经网络；

[0032] 所述的DAC电路用于将输入层的数字输入转化为模拟输入；

[0033] 所述卷积层电路与全连接层电路均采用所述基于模拟矩阵运算单元的卷积运算实 现功能并构建电路；

[0034] 所述的激活函数电路将卷积层的输出电流转换为电压，同时完成激活函数作用，包 括ReLU激活函数实现和sigmoid函数实现；

[0035] 所述的采样保持电路用于将卷积层的序列计算结果采样并保持在同一时间点；

[0036] 所述的池化层电路用于池化层的实现，包括平均池化电路和最大池化电路；

[0037] 所述的ADC电路用于将输出层的模拟输出转换为数字输出；

[0038] 具体步骤包括：

[0039] 输入特征图经过DAC电路转换为模拟信号输入；

[0040] 经过卷积层电路得到第一次卷积计算结果，然后经过激活函数电路，将计算结果从 电流转变为电压，并且进行非线性化；

[0041] 经过采样保持电路，得到一个完整的中间特征图，再经过池化电路，将模拟信号降维；

[0042] 继续经过全连接层电路进行全连接运算，最后经过激活函数电路得到模拟输出特征 图，并经ADC电路转换为数字输出。

[0043] 优选地，所述全连接层电路为一个P行Q列的的模拟矩阵运算单元，具体实现方法为：

[0044] 将输入层$[a_1\cdots a_Q]$按照每个乘法电路中模拟电压输入与所述两个电流输入端电压差值对应一个a_i其中$i\in(1\cdots Q)$，从模拟矩阵运算单元模拟电压输入端输入；

[0045] 将全连接层的P×Q个权重拆分为P个横向量，即$[w_{11},w_{21},\cdots w_{Q1}]$，$[w_{12},w_{22},\cdots w_{Q2}]$，……$[w_{1P},w_{2P},\cdots w_{QP}]$，按照一对浮栅场效应管的阈值电压差值对应 一个乘数进行存储，将其映射到模拟矩阵运算单元的P行中，完成大小为P×Q的权重 矩阵的存储；

[0046] 经过一个计算周期，通过采样保持器将所有输出的运算结果采样保持至一个时间点，得到完整的卷积结果，输出层向量[b_1,b_2,…b_p]。

[0047] 与现有技术相比，本发明具有如下的有益效果：

[0048] (1)本发明采用模拟矩阵运算单元实现存储与计算一体化进行卷积运算，模拟运算单元在存储权重的同时完成该权重的相关运算，省去了数据的往返传输；采用多脉冲 下向量依次输入和依次输出，并对输出进行采样保持，相对于冯诺依曼结构的存储运算 分隔体系，节省了90％以上的能效，另一方面模拟计算单元能够完成多位计算且模拟运 算单元之间具有高度的计算并行性，相对于数字计算单元实现，极大地减少了面积，大 大的提高了的计算密度和效率，此外对于组成模拟矩阵运算单元的浮栅场效应管阵列进 行栅极控制选择，可以选定模拟乘加单元的个数，从而快速决定计算的规模，使得卷积 运算的实用性和适用性更强。

[0049] (2)本发明构成模拟矩阵运算单元的模拟乘法电路通过阈值电压差值存储乘数，模拟电压输入与所述两个电流输入端电压差值的形式输入被乘数，以电流差值的函数读出乘积的方式实现乘法运算，由于乘法的乘数是预先存储在闪存单元中的，通过预先存 储能够多次复用的乘数，能极大的减少运算数据的读取，并且通过不同的电路拓扑结构 可拓展为向量的点积以及向量与矩阵的乘法。

[0050] (3)本发明用一对浮栅场效应管的阈值电压差值存储一个乘数，以电流差值的形式 获取乘积的方法，比仅采用一个浮栅场效应管进行存储的线性度更好，可以达到更高的精度。

[0051] (4)本发明对于差分电流检测电路的改进型设计，一方面能够使得电流输入端的电 压保持稳定，避免了现有技术中当电流改变时，电流流过负载成为负载上的电压明显会改变从而引起的计算误差，另一方面，特别的采用工作在深三极管区的浮栅场效应管或MOS管代替常用的电阻作为电流-电压转换器中的阻性负载，实现了输出电流流经作为 负载的浮栅场效应管或MOS管，可以抵消工艺参数的影响，确保在不同温度和不同工 艺下的乘法结果一致性。

[0052] (5)本发明更优异的地方在于输出电流流经作为负载的浮栅场效应管，可以通过控制调整作为负载的浮栅场效应管的阈值电压，实现对乘法结果的比例进行灵活缩放的效果。

[0053] (6)本发明成对出现的浮栅场效应管，能够有效抵消阈值电压受到的体效应的影响，从而保证乘积结果的一致性，因此乘数与被乘数都可以为正值、负值或零，扩大了 乘积使用范围，能够完成多比特运算。

[0054] (7)本发明在浮栅阵列的基础上实现存算一体化，用于卷积神经网络的推理过程，浮栅单元存储权重参数同时完成和此权重相关的乘加法运算，在此基础上完成了卷积神经网络中卷积层和全连接层的运算实现，并配合具体的池化层模块和激活函数模块电 路，实现了多层卷积神经网络，具有较好的通用性。

[0055] 关于本发明相对于现有技术，其他突出的实质性特点和显著的进步在实施例部分进 一步详细介绍。

附图说明

[0056] 通过阅读参照以下附图对非限制性实施例所作的详细描述,本发明的其它特 征、目的和优点将会变得更明显:

[0057] 图1是本发明的模拟乘法电路的结构图;

[0058] 图2a是本发明采用的一般差分电流检测电路的结构图;

[0059] 图2b是电流-电压转换器结构图;

[0060] 图2c是电压差分放大器的结构图;

[0061] 图3a是本发明采用MOS管作为阻性负载的改进型差分电流检测电路的结构图;

[0062] 图3b是本发明采用浮栅场效应管作为阻性负载的改进型差分电流检测电路的 结构图;

[0063] 图3c是本发明采用MOS管作为阻性负载的电流-电压转换器结构图;

[0064] 图3d是本发明采用浮栅场效应管作为阻性负载的电流-电压转换器结构图;

[0065] 图4是本发明采用改进型差分电流检测电路的模拟乘法电路的结构图;

[0066] 图5是本发明乘累加电路的结构图;

[0067] 图6是本发明卷积运算采用的模拟矩阵运算单元电路结构图;

[0068] 图7是本发明的带数据选择器的模拟矩阵运算单元电路结构图;

[0069] 图8是本发明的采用改进型差分电流检测电路的模拟矩阵运算单元电路结构 图;

[0070] 图9是本发明的采用改进型差分电流检测电路的带数据选择器的模拟矩阵运算单元结构图;

[0071] 图10是本发明模拟乘法方法的流程图;

[0072] 图11是本发明的一种应用在模拟矩阵运算单元中的乘法方法的流程图;

[0073] 图12是本发明的基于模拟矩阵运算单元实现2D卷积运算的示意图;

[0074] 图13是本发明的基于模拟矩阵运算单元实现2D卷积运算的原理图;

[0075] 图14是本发明的基于模拟矩阵运算单元实现一个具体的3D卷积运算的示意图;

[0076] 图15是本发明的基于模拟矩阵运算单元实现一个具体的3D卷积运算的原理图;

[0077] 图16是本发明的基于模拟矩阵运算单元实现一个具体的3D卷积运算的时序图;

[0078] 图17是本发明的基于模拟矩阵运算单元实现一般3D卷积操作的示意图;

[0079] 图18是本发明的基于模拟矩阵运算单元实现一般3D卷积操作中将3D卷积核的权重参数映射到模拟矩阵计算单元的流程图;

[0080] 图19是本发明的基于模拟矩阵运算单元实现一般3D卷积操作的时序图;

[0081] 图20是本发明的一种采用上述的基于模拟矩阵运算单元的卷积运算实现卷积神经 网络的方法流程示意图;

[0082] 图21是本发明提供的一种ReLU激活函数电路图;

[0083] 图22是本发明提供的一种sigmoid激活函数电路图;

[0084] 图23是本发明提供的一种输入特征图分别经过最大值池化和平均值池化的示意图;

[0085] 图24是本发明提供的一种输入特征图最大值池化的电路原理图;

[0086] 图25是本发明提供的一种全连接层的输入输出示意图;

[0087] 图26是本发明提供的基于模拟矩阵运算单元进行全连接运算的示意图。

具体实施方式

[0088] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实 施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下 所获得的所有其他实施例，都属于本发明保护的范围。

[0089] 参照说明书附图1-26，对本发明一种基于模拟矩阵运算单元的卷积运算及其应用进行具体实施例描述：

[0090] 实施例1

[0091] 如图1所示，为本实施例中构成模拟矩阵运算单元的单个模拟乘法电路，包括一对浮栅场效应管M_1、M_2和一个差分电流检测电路；

[0092] 其中M_1和M_2共栅极并接入电压源，共漏极或共源极并接入模拟电压输入，差分电流检测电路包括两个电流输入端和一个输出端，所述两个电流输入端对应接入M_1和M_2各自源极或漏极；在这里说明的是如果M_1和M_2共漏极接入模拟电压输入，则两个电流 输入端对应接入M_1和M_2各自源极，如果M_1和M_2共源极接入模拟电压输入，则两个 电流输入端对应接入M_1和M_2各自漏极，这是由于浮栅场效应管的源漏极可以互换的结 果；

[0093] 本实施例中采用的差分电流检测电路能保持两个电流输入端电压不变，所述输出端 结果为两个电流输入端输入电流差值的函数；

[0094] 如图2a所示是满足本实施例功能的一般差分电流检测电路：

[0095] 所述差分电流检测电路包括两个电流-电压转换器和一个电压差分放大器；所述电 流-电压转换器由一个运算放大器和一个阻性负载构成，运算放大器的负极输入端为所述电流-电压转换器的电流输入端口，运算放大器的输出端为所述电流-电压转换器的电压输出端口，运算放大器的正极输入端外加直流电平V_b，所述阻性负载连接在运算放大 器的负极输入端与运算放大器的输出端之间；所述电压差分放大器由一个运算放大器、两个阻值相同的电阻R_1、R_3和两个阻值相同的电阻R_2、R_4构成，所述电压差分放大器 的两个电压输入端口分别通过电阻R_1、R_3连接到运算放大器的正、负极输入端，运算放 大器的正极输入端通过电阻R_2连接到直流电平V_b，运算放大器的负极输入端通过电阻 R_4连接到运算放大器的输出端，运算放大器的输出端为所述电压差分放大器的电压输出 端口；

[0096] 具体的，图2a的差分电流检测电路是由图2b的电流-电压转换器和图2c的电压差分放大器构成的。图2b的电流-电压转换器的输入电流为I_{in}，通过负载R转换为输出电 压V_{out}，输出与输入有如下关系：$V_{out}=V_b+I_{in}R$，实现了电流-电压的转换；图2c的电 压差分放大器的输出与输入有如下关系：$V_{out}=V_b+\frac{R_2}{R_1}(V_{in1}-V_{in2})$，实现了差分电压放大。

[0097] 参考图10，在上述模拟乘法电路的实现乘法方法，包括以下步骤：

[0098] 步骤S_1、对两个浮栅场效应管M_1、M_2中的浮置栅极执行擦除和写操作，将乘数以 所述一对浮栅场效应管的阈值电压差值的形式进行存储；

[0099] 步骤S_2、对两个浮栅场效应管源极、漏极和选栅极执行电压施加，以所述模拟电压输入与所述两个电流输入端电压差值的形式输入被乘数；

[0100] 步骤S_3、通过差分电流检测电路输出基于所述一对浮栅场效应管输出电流差值的函 数作为所述被乘数与乘数的乘积结果。

[0101] 具体说明上述乘法方法：

[0102] 以图1所述电路结构图为例，对于M_1和M_2以读取操作进行乘法运算时，M_1和M_2须确保工作在三极管区，而栅源极电压V_{GS}、漏源电压V_{DS}均相同，在此处$V_{DS}=V_{in}-V_b$，读出 M_1和M_2的输出电流$I_{D1,2}$为 $I_{D1,2}=\mu C_{ox}\frac{W}{L}\left[(V_{GS}-V_{TH1,2})V_{DS}-\frac{1}{2}V_{DS}^2\right]$，$M_1$和$M_2$的电流差值 $\Delta I_D=I_{D1}-I_{D2}=\mu C_{ox}\frac{W}{L}(V_{TH2}-V_{TH1})V_{DS}$可以表示为阈值电压差值$V_{TH2}-V_{TH1}$与漏源电压 V_{DS}的乘积，以阈值电压差值$V_{TH1}-V_{TH2}$为乘数，漏源电压V_{DS}为被乘数即可实现乘法。

[0103] 在本实施例中用一对浮栅场效应管的阈值电压差值存储一个乘数，以电流差值的形 式获取乘积的方法，乘积结果有很好的线性度，可以达到较高的精度。

[0104] 为了使M_1和M_2都工作在三极管区，需要满足M_1和M_2的栅源电压大于两者阈值电压的最大值也即$V_{GS}>\max\{V_{TH1},V_{TH2}\}$，漏源电压小于栅源电压分别与两者阈值电压差值 的最小值，也即$V_{DS}<\min\{V_{GS}-V_{TH1},V_{GS}-V_{TH2}\}$。

[0105] 在V_{GS}固定的情况下，需要对阈值电压V_{TH}的动态范围与漏源电压V_{DS}的动态范围进行 折中选择。对于乘数的存储方式，所述M_1和M_2的栅源电压V_{GS}为固定值，漏源电压V_{DS}取值范围为$[0,V_{DS(max)}]$，阈值电压V_{TH1}、V_{TH2}的取值范围为$[0,V_{TH(max)}]$：

[0106] 当$V_{TH1}-V_{TH2}=\Delta V_{TH}>0$时，选择$V_{TH1}=V_{TH(max)}$，$V_{TH2}=V_{TH(max)}-\Delta V_{TH}$，当 $V_{TH1}-V_{TH2}=-\Delta V_{TH}<0$时，选择$V_{TH2}=V_{TH(max)}$，$V_{TH1}=V_{TH(max)}-\Delta V_{TH}$；这种选择是为了选 择较大的阈值电压可以进一步减小浮栅场效应管的电流以减小功耗。

[0107] 场效应管的体效应：阈值电压会受源极-衬底的电压V_{SB}影响，一般施加的衬底电压V_B是固定的，p衬底接地，n衬底接最高电位，所以只用看V_S当被乘数$V_{in}-V_b>0$时，$V_S=V_b$是恒定值；当被乘数$V_{in}-V_b<0$时，$V_S=V_{in}$是随输入改变的，所以阈值电压 也随输入改变。具体的在本实施例中，如图2所示输入的被乘数为漏源电压即$V_{in}-V_b$。当$V_{in}>V_b$时，电流由电压输入端流向差分电流检测电路的输入端口，两个闪存单元的 上端为漏极，下端为源极；当$V_{in}<V_b$时，电流由差分电流检测电路的输入端口流向电 压输入端，两个闪存单元的上端为源极，下端为漏极，此时由于源极的电位由外部输入， 闪存单元的阈值电压受到体效应的影响，变化量为$\Delta V_{TH}=\gamma(\sqrt{|2\Phi_F+V_{SB}|}-\sqrt{|2\Phi_F|})$，由 于两个闪存单元的的源极-衬底电压$V_{SB}$始终相同，阈值电压受到体效应影响的变化量 $\Delta V_{TH1,2}$相同，因此存储的阈值电压差值$V_{TH1}-V_{TH2}$不变，乘积结果不受影响。因此，被 乘数可以为正值、负值或零。

[0108] 在本实施例中，可以继续对差分电流检测电路进行进一步改进，如图3a和3b所示，主要是阻性负载采用浮栅场效应管或MOS管，所述浮栅场效应管或MOS管的栅极外加 栅极电压，漏极连接到所述运算放大器的输出端，源极连接到所述运算放大器的负极输 入端；

[0109] 为了保证浮栅场效应管或MOS管工作在深三极管区，所述浮栅场效应管或MOS管的漏源电压V_{DS}、栅源电压V_{GS}和阈值电压V_{TH}满足：$V_{DS}<<2(V_{GS}-V_{TH})$。

[0110] 更进一步具体的说明，参照图3c与图3d中的电流-电压转换器分别采用MOS管或浮栅场效应管作为负载，工作在深三极管区的阻抗为 $R_{M_R}=\frac{1}{\mu C_{ox}\frac{W}{L}(V_{GS}-V_{TH})}$。

[0111] 在此改进的差分电流检测电路中输出电流流经作为负载的浮栅场效应管或MOS管，可以抵消工艺参数的影响，确保在不同温度和不同工艺角下的乘法结果一致性。

[0112] 继续参考图4，乘数以浮栅场效应管的阈值电压差值V_{TH1}-V_{TH2}存储，可以为正值、负值或零，被乘数以输入电压值V_{in}输入。闪存单元M1、M2的电流为$I_{D1,2}=\mu C_{ox}\frac{W}{L}\left[(V_G-V_{in}-V_{TH1,2})(V_b-V_{in})-\frac{1}{2}(V_b-V_{in})^2\right]$，通过电流-电压转换器转换为电压$V_{1,2}=V_b+\frac{(V_G-V_{in}-V_{TH1,2})(V_b-V_{in})-\frac{1}{2}(V_b-V_{in})^2}{V_{GS(M_R)}-V_{TH(MR)}}$，最后通过电压差分放大器得到输出$V_{out}=V_b+\frac{R_2}{R_1}\frac{(V_{TH2}-V_{TH1})(V_b-V_{in})}{V_{GS(M_R)}-V_{TH(MR)}}=V_b+K(V_{TH2}-V_{TH1})(V_b-V_{in})$，最终的乘法结果为$V_{out}$-$V_b$=K($V_{TH2}$-$V_{TH1}$)($V_b$-$V_{in}$)，系数$K=\frac{R_2}{R_1}\frac{1}{V_{GS(M_R)}-V_{TH(MR)}}$与工艺参数μ$C_{ox}$无关，且 电阻为比值的形式。

[0113] 在此改进型差分电流检测电路中，输出电流流经作为负载的浮栅场效应管，可以通 过控制调整作为负载的闪存单元的阈值电压，实现对乘法结果的比例进行灵活缩放的效 果。

[0114] 参考图5，本实施例中直接构成模拟矩阵运算单元一种乘累加电路，包括若干个模拟乘法电路，所述若干个模拟乘法电路共栅极且共同一个差分电流检测电路，所述每个 模拟乘法电路中该对浮栅场效应管的共漏极或共源极分别接入对应的输入电压信号；

[0115] 所述差分电流检测电路的输出端结果为每个模拟乘法电路输出的基于该对浮栅场 效应管输出电流差值的函数之和。

[0116] 在本实施例中差分电流检测电路可以获取多对所述浮栅场效应管电流差值之和，或 者，可以获取多对所述浮栅场效应管中的第一个浮栅场效应管电流之和，以及多对所述浮栅场效应管中的第二个浮栅场效应管电流之和，再得到其差值，以实现乘累加运算的 效果。

[0117] 参考图6是本实施例中卷积运算采用的模拟矩阵运算单元，所述模拟矩阵运算单元为一个P行Q列的模块电路，所述每一行均为一个包括Q个模拟乘法电路的乘累加电 路；

[0118] 所述每列浮栅场效应管的共漏极或共源极接入同一的输入电压信号。

[0119] 具体而言本实施例中P行Q列的模块电路，参照图11中的乘法流程图，按照一对浮栅 场效应管的阈值电压差值对应一个乘数进行存储，完成大小为P×Q的乘数矩阵的存储$\begin{pmatrix} b_{11} & \cdots & b_{1Q} \\ \vdots & \ddots & \vdots \\ b_{P1} & \cdots & b_{PQ} \end{pmatrix}$；每列浮栅场效应管的一端源或漏极相连构成位线(BL)信号，以每个 乘法电路中模拟电压输入与所述两个电流输入端电压差值对应一个被乘数，完成元素个 数为Q的横向被乘数向量的输入($a_1\cdots a_Q$)；模块电路中中Q行中每行的栅极接入同一个字 线(WL)信号，另一端源或漏极每行相连构成两条源/漏极线(SDL)，接到一个差分电 流检测电路，产生一个乘累加输出，所有的Q个输出可表示为输出向量

$(c_1\cdots c_Q)=K(a_1\cdots a_Q)\begin{pmatrix} b_{11} & \cdots & b_{1Q} \\ \vdots & \ddots & \vdots \\ b_{P1} & \cdots & b_{PQ} \end{pmatrix}^T$，实现了向量与矩阵的乘法。

[0120] 参照图8，是本实施例中模拟矩阵运算单元也可以采用改进型差分电流检测电路的 实现的电路示意图，其差分电流检测电路为采用浮栅场效应管作为阻性负载的改进型差分电流检测电路，实现向量与矩阵的乘法：$(c_1\cdots c_Q)=K(a_1\cdots a_Q)\begin{pmatrix} b_{11} & \cdots & b_{1Q} \\ \vdots & \ddots & \vdots \\ b_{P1} & \cdots & b_{PQ} \end{pmatrix}^T$，被 乘数$a_j=V_{in(j)}-V_b$，乘数$b_{ij}=V_{TH(i,j)}-V'_{TH(i,j)}$，系数$K=\frac{R_2}{R_1}\frac{1}{V_{GS(M_R)}-V_{TH(MR)}}$，乘法结 果$c_i=V_{out(i)}-V_b$。

[0121] 参照图7，本实施例中进一步改进的模拟矩阵运算单元，包括第一数据选择器和第二数据选择器，所述第一数据选择器设置在每列浮栅场效应管的共漏极或共源极与其对应的输入电压信号之间，所述第二数据选择器设置在每行浮栅场效应管的共源极或共漏极与差分电流检测电路的电流输入端之间；

[0122] 所述第一、二数据选择器组合用于从P行Q列的模块电路任意选取X行、Y列单元形成 新的工作模块；以减少电路中输入信号、差分电流检测电路与输出信号的数量。

[0123] 参照图9，本实施例中进一步改进的模拟矩阵运算单元电路结构示意图，其差分电流检测电路为采用浮栅场效应管作为阻性负载的改进型差分电流检测电路。

[0124] 本实施例中卷积运算包括以下步骤：

[0125] 将卷积核经过转换拼接成P个长度为Q的横向量，并映射到矩阵$\begin{pmatrix} b_{11} & \cdots & b_{1Q} \\ \vdots & \ddots & \vdots \\ b_{P1} & \cdots & b_{PQ} \end{pmatrix}$中；

[0126] 将输入特征图按照滑动窗口进行切分转换成N个长度为Q的横向量；

[0127] 在N个脉冲波下，依次将N个长度为Q的横向量映射到向量$(a_1\cdots a_Q)$中，矩阵运 算单元的模拟输出端口按照脉冲波序列，输出N个长度为P的运算结果；

[0128] 通过采样保持器将所有输出的运算结果采样保持至一个时间点，得到完整的卷积结 果。

[0129] 下面以具体维度的卷积计算解释实施例1中的如何基于上述模拟矩阵运算单元实现 卷积方法：

[0130] 实施例2

[0131] 参考图12，是本实施例的一个基于模拟矩阵运算单元实现2D卷积运算的示意图：其中输入特征图的维度为5×5，输出特征图的维度为3×3，卷积核的维度为3×3，未 考虑偏置bias。

[0132] 图13是该2D卷积用模拟矩阵运算单元实现的原理图；将大小为3×3卷积核转换为横向量$[w_{11},w_{12},w_{13},w_{21},w_{22},w_{23},w_{31},w_{32},w_{33}]$将其映射到模拟矩阵运算单元的一行；将 输入特征图按照滑动窗口转换为9个横向量，即$[a_{11},a_{12},a_{13},a_{21},a_{22},a_{23},a_{31},a_{32},a_{33}]$，$[a_{12},$

$a_{13},a_{14},a_{22},a_{23},a_{24},a_{32},a_{33},a_{34}]$,……,$[a_{33},a_{34},a_{35},a_{43},a_{44},a_{45},a_{53},a_{54},a_{55}]$,将其转换 为模拟电压,然后在9个脉冲波下从模拟输入端口并行序列依次输入;在模拟输出端口 依次输出9个运算结果,$[b_{11},b_{12},b_{13},b_{21},b_{22},b_{23},b_{31},b_{32},b_{33}]$。通过采样保持器,将所有运 算结果采样保持至一个时间点,得到一个3×3的输出特征图。1个脉冲进行一次滑窗运 算,9个脉冲构成一个完整的卷积周期,进行一次完整的卷积计算,输出一个完整2D 卷积结果;

[0133] 实施例3

[0134] 参考图14,是本实施例基于模拟矩阵运算单元实现一个具体的3D卷积运算的示意图:

[0135] 其中输入特征图维度为3×3×3,输出特征图维度为3×3×2,采用了2个1×1× 3的卷积核,2个1×1×1的偏置bias。

[0136] 参考图15,是该3D卷积用模拟矩阵运算单元实现的原理图,将2个大小为1×1 ×3卷积核和2个1×1×3的偏置bias转换为2个横向量$[w_{11},w_{12},w_{13},b_1]$,$[w_{21},w_{22},w_{23},b_2]$, 将其映射到模拟矩阵运算单元的两行;将输入特征图按照滑动窗口转换为9个横向量, 即$[a_{111},a_{112},a_{113},1]$,$[a_{121},a_{122},a_{123},1]$,……,$[a_{331},a_{332},a_{333},1]$。其转换为模拟电压,然后 在9个脉冲波下从模拟输入端口并行序列依次输入;在模拟输出端口依次输出9个运算 结果$[b_{111},b_{112}]$,$[b_{121},b_{122}]$,……,$[b_{331},b_{332}]$,通过采样保持器,将所有运算结果采样保 持至一个时间点,得到一个3×3×2的输出特征图。一个脉冲进行1次滑窗运算,9个 脉冲构成一个完整的卷积周期,进行一次完整的卷积计算,输出一个完整3D卷积结果;

[0137] 参考图16是该3D卷积运算的时序图。

[0138] 实施例4

[0139] 参照图17-19,本实施例是基于模拟矩阵运算单元的一般3D卷积操作;

[0140] 图17是该操作的示意图,其中输入特征图的维度是$n_H^{[l-1]}\times n_W^{[l-1]}\times n_c^{[l-1]}$(即第1-1层的特征图维度(长×宽×通道)),输出特征图的维度是$n_H^{[l]}\times n_W^{[l]}\times n_c^{[l]}$(即第1层的特征图 维度(长×宽×通道)),卷积操作使用了$n_c^{[l]}$个维度为$f^{[l]}\times f^{[l]}\times n_c^{[l-1]}$(即第1层的3D 卷积核维度(长×宽×通道))的卷积核;

[0141] 图18是将3D卷积核的权重参数映射到模拟矩阵计算单元的流程图;每个3D卷积核的维度为$f^{[l]}\times f^{[l]}\times n_c^{[l-1]}$,将其切片成$n_c^{[l-1]}$个大小为$f^{[1]}\times f^{[1]}$的2D卷积核,然后将这 些2D卷积核转换为横向量,拼接成一个横向量,即 $[w_{111},w_{121},\cdots,w_{ff1},w_{112},\cdots,w_{ff2},\cdots,w_{ffn}]$,向量长度为$f^{[l]}\times f^{[l]}\times n_c^{[l-1]}$;一共有$n_c^{[l]}$个 大小为$f^{[l]}\times f^{[l]}\times n_c^{[l-1]}$的3D卷积核,将其全部切片展开为$n_c^{[l]}$个长度为$f^{[l]}\times f^{[l]}\times n_c^{[l-1]}$的 横向量,然后将它们拼接为一个维度为$n_c^{l}\times(f^{[l]}\times f^{[l]}\times n_c^{[l-1]})$的权重矩阵,将其映射到矩 阵运算模块,具体为:权重矩阵第一行映射模拟矩阵计算单元第一行,矩阵第二行映射 模拟矩阵计算单元第二行,矩阵第三行映射模拟矩阵计算单元第三行……

[0142] 图19是3D卷积操作进行运算的时序图;进行一次完整的卷积操作需要$n_H^{[l]}\times n_W^{[l]}$个脉冲周期,将输入特征图按照滑动窗口进行切分,

[0143] 可以转化为$n_H^{[l]}\times n_W^{[l]}$个横向量,每个横向量的长度为$f^{[l]}\times f^{[l]}\times n_c^{[l-1]}$,即 $[a_{11},$

$a_{12},\cdots,a_{1f},a_{21},\cdots,a_{2f},\cdots,a_{ff}]\cdots[a_{12},a_{13},\cdots,a_{ff+1},a_{22},\cdots,a_{2f+1},\cdots,a_{f+1f+1}]$。将其通过DAC转换为模拟电压，然后在$n_H^{[l]}\times n_W^{[l]}$个脉冲波下从模拟输入端口并行序列依 次输入；矩阵运算单元的模拟输出端口按照脉冲波序列输出$n_H^{[l]}\times n_W^{[l]}$个长度为$n_c^{[l]}$的运算 结果，通过采样保持器，将其输出保持在一个时间点，于是得到一个大小为$n_H^{[l]}\times n_W^{[l]}\times n_c^{[l]}$的3D输出矩阵；一个脉冲进行一次滑窗运算，$n_H^{[l]}\times n_W^{[l]}$个脉冲构成一卷积运算周期，进 行一次完整的卷积计算，输出一个完整3D卷积结果。

[0144] 实施例5

[0145] 参照图20，本实施例提供一种采用上述的基于模拟矩阵运算单元的卷积运算实现卷 积神经网络的方法，实现了一个输入层，卷积层，池化层，全连接层，输出层共5层得 卷积神经网络，具体包括DAC电路、卷积层电路、激活函数电路、采样保持电路、池化 层电路、全连接层电路和ADC电路，各个电路间的级联构成卷积神经网络；

[0146] 所述的DAC电路用于将输入层的数字输入转化为模拟输入；

[0147] 所述卷积层电路与全连接层电路均采用所述基于模拟矩阵运算单元的卷积运算实 现功能并构建电路；

[0148] 所述的激活函数电路将卷积层的输出电流转换为电压，同时完成激活函数作用；

[0149] 所述的采样保持电路用于将卷积层的序列计算结果采样并保持在同一时间点；

[0150] 所述的池化层电路用于池化层的实现，包括平均池化电路和最大池化电路；

[0151] 所述的ADC电路用于将输出层的模拟输出转换为数字输出；

[0152] 具体步骤包括：

[0153] 输入特征图经过DAC电路转换为模拟信号输入；

[0154] 经过卷积层电路得到第一次卷积计算结果，然后经过激活函数电路，将计算结果从 电流转变为电压，并且进行非线性化；

[0155] 经过采样保持电路，得到一个完整的中间特征图，再经过池化电路，将模拟信号降维；

[0156] 继续经过全连接层电路进行全连接运算，最后经过激活函数电路得到模拟输出特征 图，并经ADC电路转换为数字输出。

[0157] 在本实施例中激活函数可以采用ReLU激活函数或者sigmoid激活函数，具体的：

[0158] 参考图21，为满足本实施例的一种ReLU激活函数电路图，该ReLU激活函数的电路包 括：电流电压转换器，反相器，电压限幅器；电流电压转换器用于将来自矩阵计算单元的模拟输出端口的电流转换为电压；反相器起到电压反向缓冲的作用；电压限幅器采用 二极管方式，将大于0的电压输出，小于0的电压保持为0；输入电流与输出电压的关系 为$V_{\text{out}}=\begin{cases}R_1\bullet I_{in} & I_{in}>0\\ 0 & I_{in}<0\end{cases}$。

[0159] 参考图22，为满足本实施例的一种sigmoid激活函数电路图，该sigmoid激活函数的 电路包括：带偏置的电流电压转换器，反相器，电压限幅器；电流电压转换器用于将来自矩阵计算单元的模拟输出端口的电流转换为电压，在电流电压转换器的负极性端串联电阻并提供偏置电压；反相器起到电压反向缓冲的作用；电压限幅器将输出电压限制在 一

定的电压范围内，输入电流与输出电压的关系为

[0160] $$V_{\text{out}}=\begin{cases}0 & I_{in}<-\dfrac{V_{bias}}{R_2}\\ R_1\bullet I_{in}+\dfrac{V_{\text{bias}}\bullet R_1}{R_2} & -\dfrac{V_{bias}}{R_2}<I_{in}<\dfrac{V_{bias}}{R_2}\\ V_{DD} & I_{in}>\dfrac{V_{bias}}{R_2}\end{cases};$$

[0161] 本实施例中还提供了一个卷积神经网络的池化层电路池化操作的示例，包括平均值 池化和最大值池化具体的：

[0162] 参考图23是一个维度为4×4的输入特征图分别经过最大值池化和平均值池化的示 意图；

[0163] 参考图24是最大值池化的电路原理图，包括配置单元和一些平均池化单元和最大池 化单元。配置单元用来根据池化降维的大小配置平均池化单元和最大池化单元的个数，一个池化单元(平均或者最大)能够进行一个池化窗口的操作，就是将多个模拟信号取 最大值或者平均值输出。最大池化单元由模拟信号比较器和模拟多路复用器构成，模拟 信号比较器比较池化窗口的多路模拟输入信号的大小，找出最大模拟输入信号，然后通 过模拟多路复用器选通最大模拟输入信号，过滤其余模拟信号；平均池化单元由模拟反 向加法器和一个电压反向器串联组成，池化窗口的多路模拟输入信号从$v_i^1,v_i^2,\cdots v_i^n$输入， 池化结果从v_0^1输出。当R1＝R2＝…＝R_n＝nR_f时，$V_0=\dfrac{v_i^1+v_i^2+\cdots+v_i^n}{n}$，起到一个平均 模拟输入电压的作用。

[0164] 本实施例中给出一个全连接层电路均采用所述基于模拟矩阵运算单元的卷积运算 实现的示例，具体的：

[0165] 参考图25是一个全连接层的示意图，输入层为[$a_1\cdots a_Q$]，输出层为[$b_1,b_2,\cdots b_P$]，共有P×Q个权重参数，未考虑偏置；

[0166] 参考图26，是用模拟矩阵运算单元进行全连接运算的示意图；将输入层[$a_1\cdots a_Q$]按照每个乘法电路中模拟电压输入与所述两个电流输入端电压差值对应一个a_i其中 i∈(1,2,…Q)，从模拟矩阵运算单元模拟电压输入端输入；

[0167] 将全连接层的P×Q个权重拆分为P个横向量，即[$w_{11},w_{21},\cdots w_{Q1}$]， [$w_{12},w_{22},\cdots w_{Q2}$]，……[$w_{1P},w_{2P},\cdots w_{QP}$]，按照一对浮栅场效应管的阈值电压差值对应 一个乘数进行存储，将其映射到模拟矩阵运算单元的P行中，完成大小为P×Q的权重 矩阵的存储；

[0168] 经过一个计算周期，通过采样保持器将所有输出的运算结果采样保持至一个时间点，得到完整的卷积结果，输出层向量[$b_1,b_2,\cdots b_P$]。

[0169] 本发明的一种模拟乘法电路、模拟乘法方法及其应用能通过以乘数预先存储、被乘 数在运算时输入的方式实现两个数的乘法、两个向量的点积以及向量与矩阵的乘法，运算速度快、电路功耗低、方法简单易行，具有较高的实用价值和广泛的应用前景。

[0170] 对于本领域技术人员而言，显然本发明不限于上述示范性实施例的细节，而且 在不背离本发明的精神或基本特征的情况下，能够以其他的具体形式实现本发明。因此，无

论从哪一点来看，均应将实施例看作是示范性的，而且是非限制性的，本 发明的范围由所附权利要求而不是上述说明限定，因此旨在将落在权利要求的等同 要件的含义和范围内的所有变化囊括在本发明内。不应将权利要求中的任何附图标 记视为限制所涉及的权利要求。

[0171] 此外，应当理解，虽然本说明书按照实施方式加以描述，但并非每个实施方式 仅包含一个独立的技术方案，说明书的这种叙述方式仅仅是为清楚起见，本领域技 术人员应当将说明书作为一个整体，各实施例中的技术方案也可以经适当组合，形 成本领域技术人员可以理解的其他实施方式。

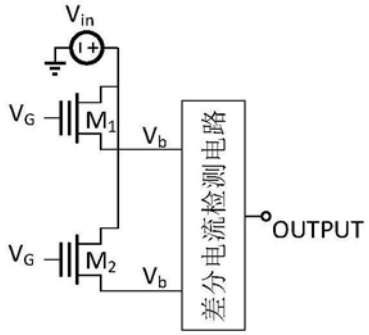


图1

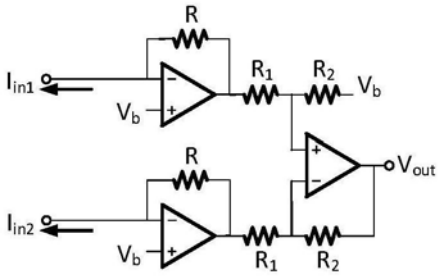


图2a

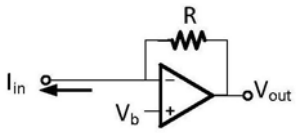


图2b

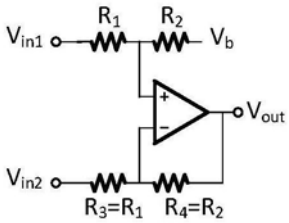


图2c

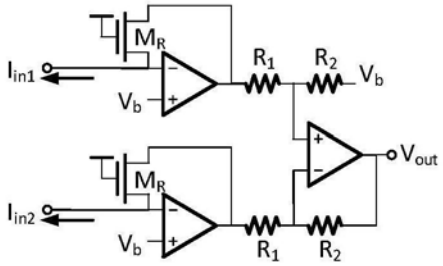


图3a

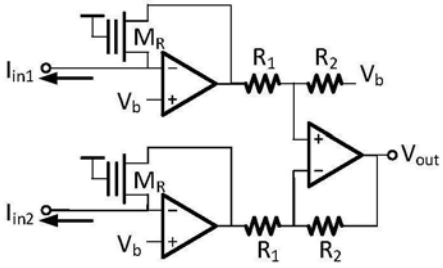


图3b

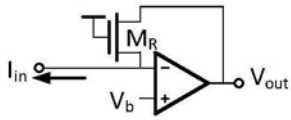


图3c

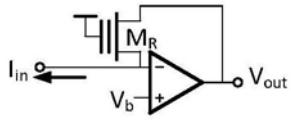


图3d

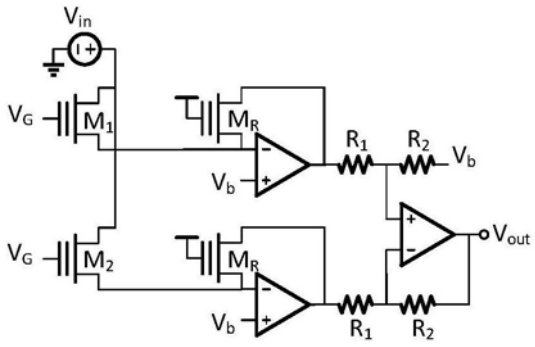

Vin
VG
M1
MR
R1
R2
Vb
Vout
VG
M2
MR
Vb
R1
R2

图4

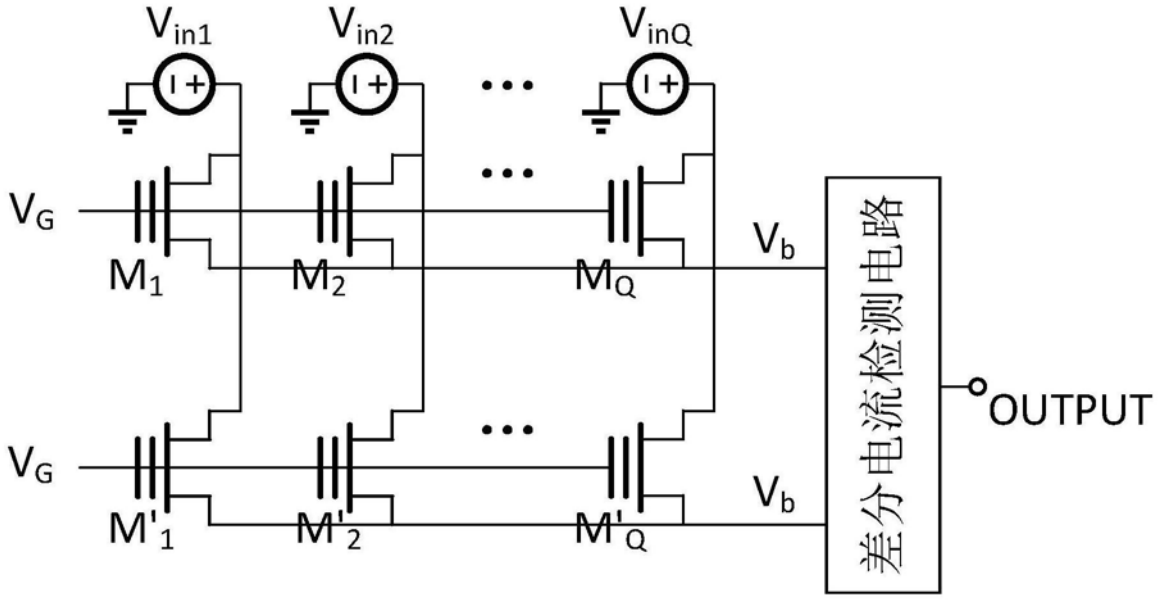

Vin1
Vin2
VinQ
VG
M1
M2
MQ
Vb
差分电流检测电路
OUTPUT
VG
M'1
M'2
M'Q
Vb

图5

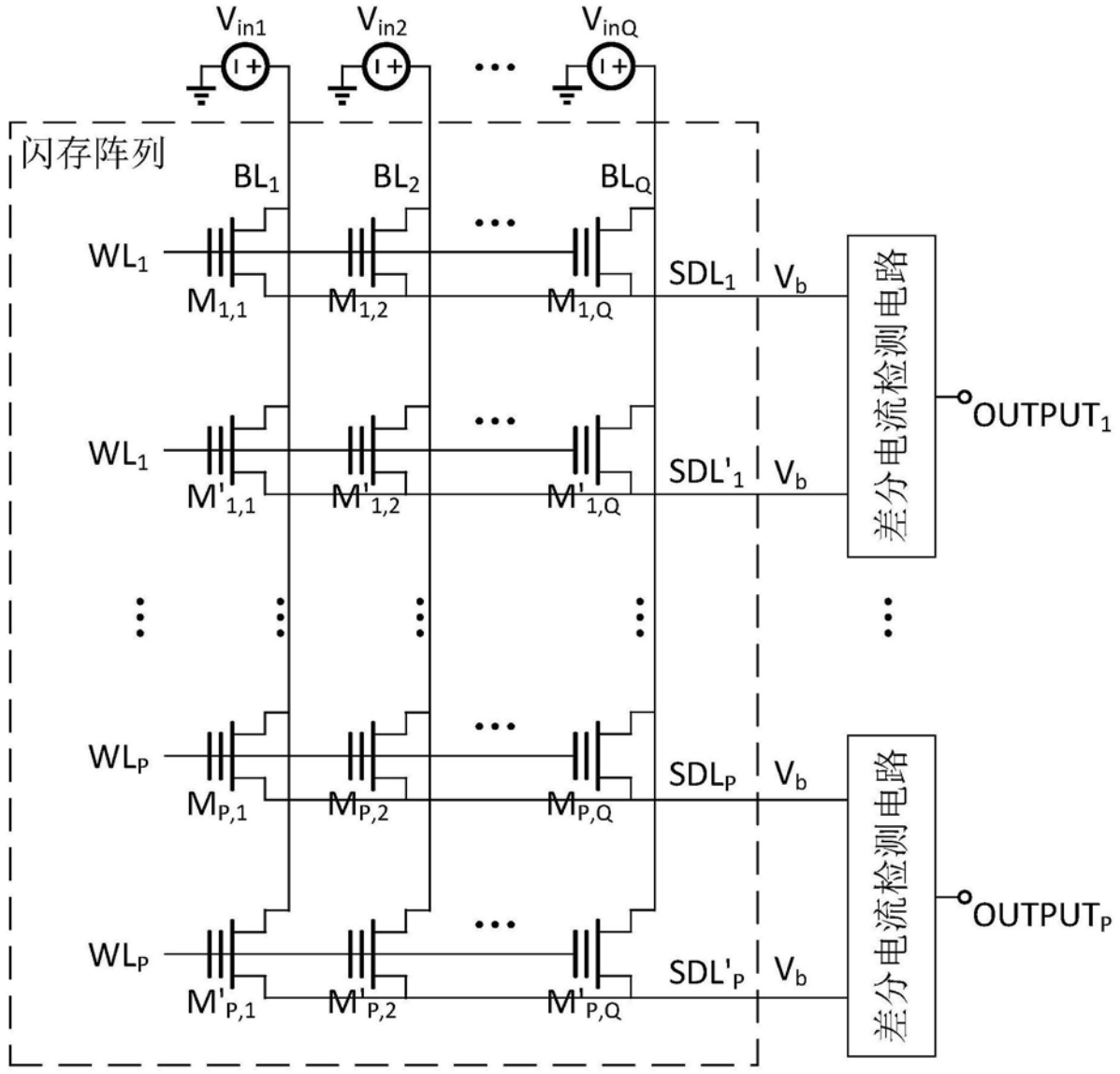

V_{in1}
V_{in2}
V_{inQ}
闪存阵列
BL_1
BL_2
BL_Q
WL_1
$M_{1,1}$
$M_{1,2}$
$M_{1,Q}$
SDL_1
V_b
WL_1
$M'_{1,1}$
$M'_{1,2}$
$M'_{1,Q}$
SDL'_1
V_b
差分电流检测电路
$OUTPUT_1$
WL_P
$M_{P,1}$
$M_{P,2}$
$M_{P,Q}$
SDL_P
V_b
WL_P
$M'_{P,1}$
$M'_{P,2}$
$M'_{P,Q}$
SDL'_P
V_b
差分电流检测电路
$OUTPUT_P$

图6

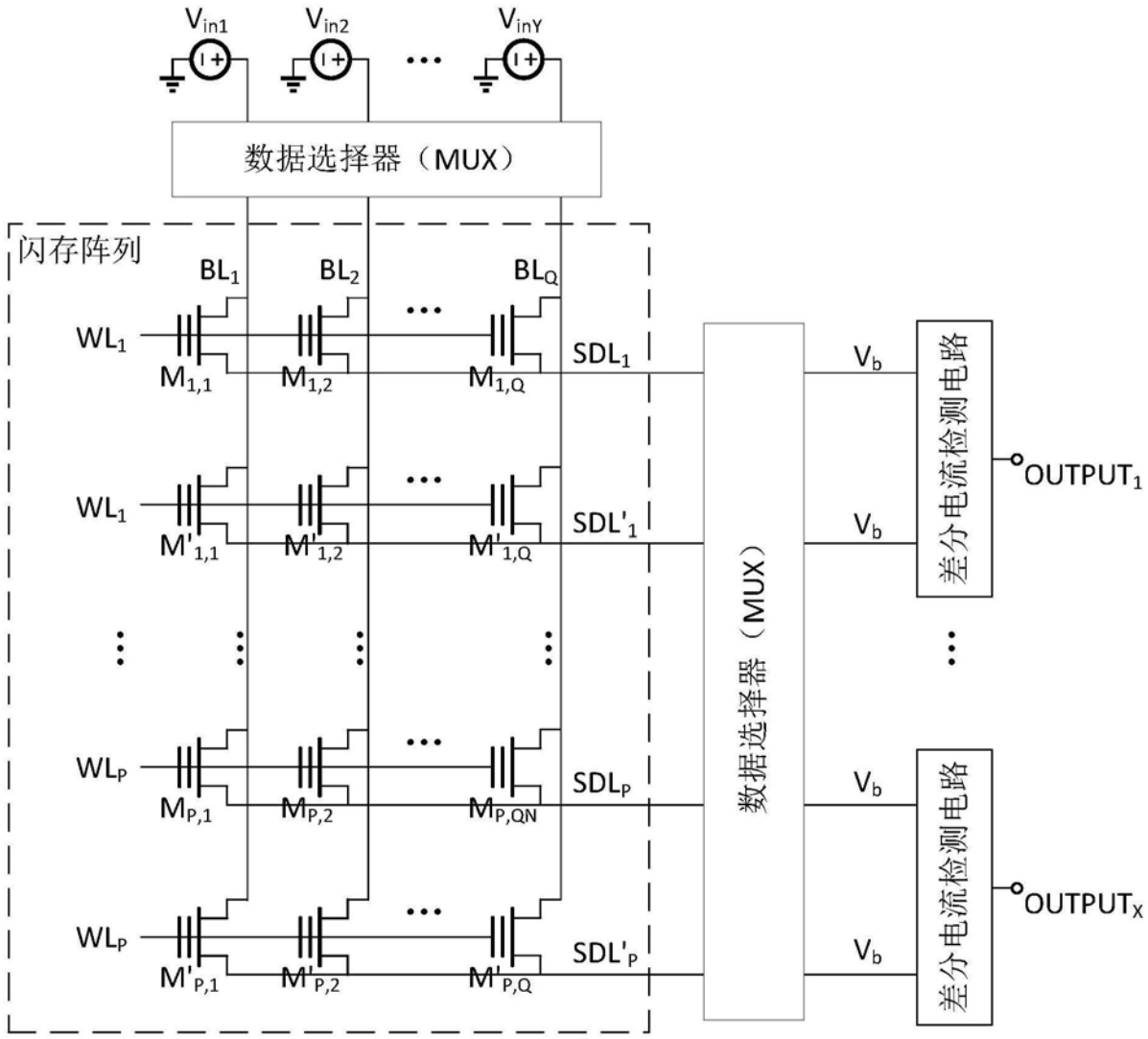

V_{in1}
V_{in2}
V_{inY}
数据选择器（MUX）
闪存阵列
BL_1
BL_2
BL_Q
WL_1
WL_P
$M_{1,1}$
$M_{1,2}$
$M_{1,Q}$
$M_{P,1}$
$M_{P,2}$
$M_{P,QN}$
$M'_{1,1}$
$M'_{1,2}$
$M'_{1,Q}$
$M'_{P,1}$
$M'_{P,2}$
$M'_{P,Q}$
SDL_1
SDL_P
SDL'_1
SDL'_P
数据选择器（MUX）
V_b
差分电流检测电路
$OUTPUT_1$
$OUTPUT_X$

图7

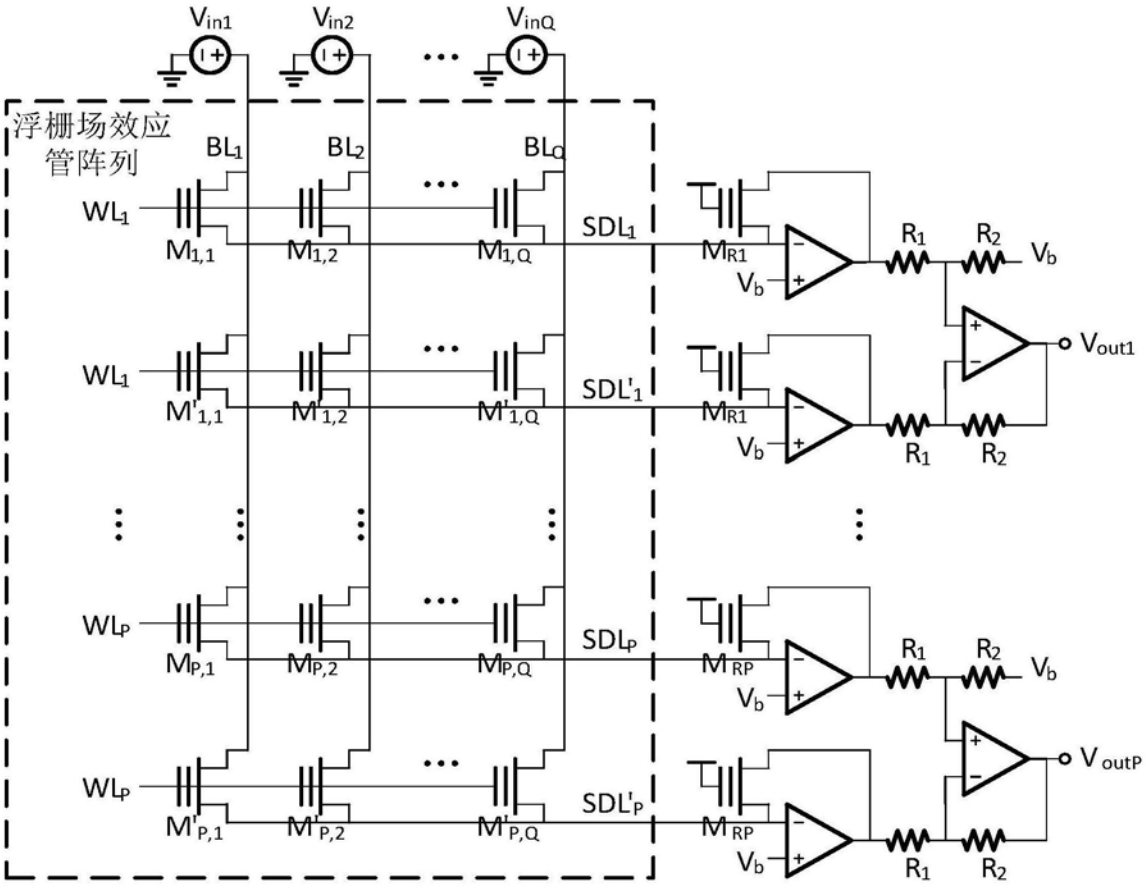

Vin1
Vin2
VinQ
浮栅场效应
管阵列
BL1
BL2
BLQ
WL1
WLP
M1,1
M1,2
M1,Q
M'1,1
M'1,2
M'1,Q
MP,1
MP,2
MP,Q
M'P,1
M'P,2
M'P,Q
SDL1
SDL'1
SDLP
SDL'P
MR1
MRP
Vb
R1
R2
Vout1
VoutP

图8

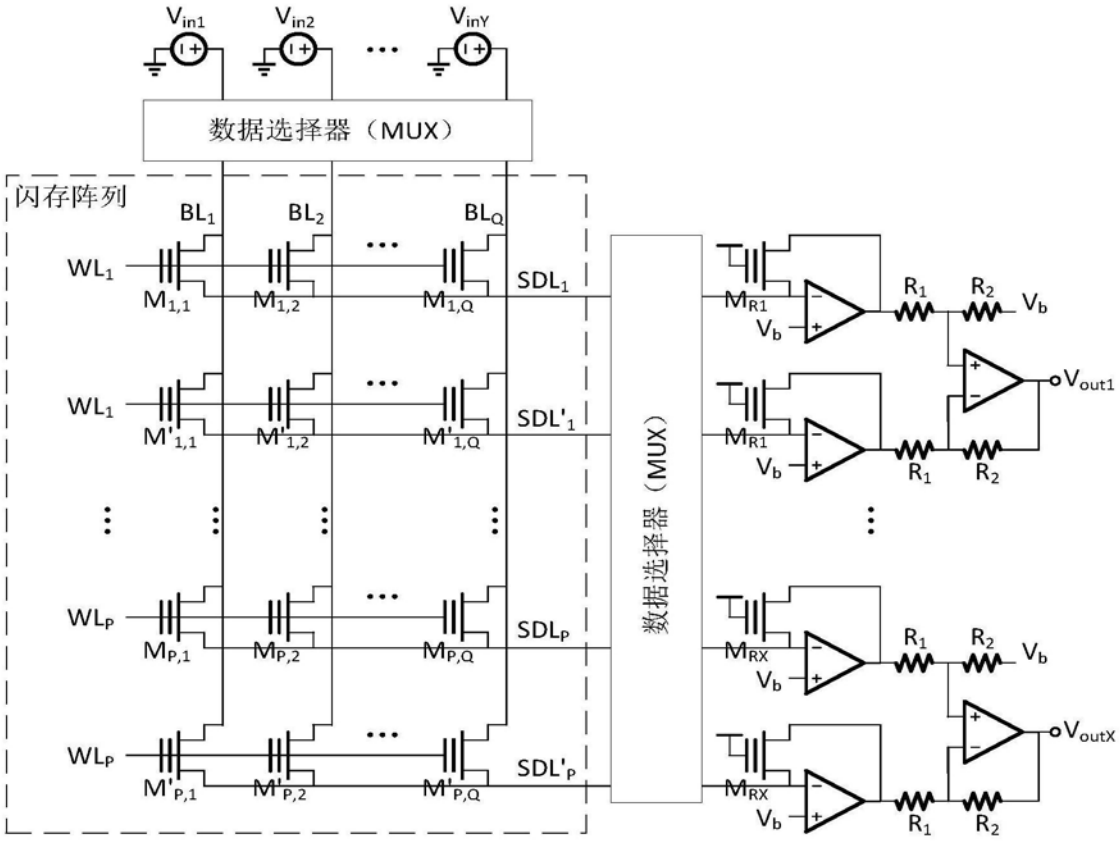

V_{in1}
V_{in2}
V_{inY}
数据选择器（MUX）
闪存阵列
BL_1
BL_2
BL_Q
WL_1
WL_P
$M_{1,1}$
$M_{1,2}$
$M_{1,Q}$
$M_{P,1}$
$M_{P,2}$
$M_{P,Q}$
$M'_{1,1}$
$M'_{1,2}$
$M'_{1,Q}$
$M'_{P,1}$
$M'_{P,2}$
$M'_{P,Q}$
SDL_1
SDL_P
SDL'_1
SDL'_P
数据选择器（MUX）
M_{R1}
M_{RX}
R_1
R_2
V_b
V_{out1}
V_{outX}

图9

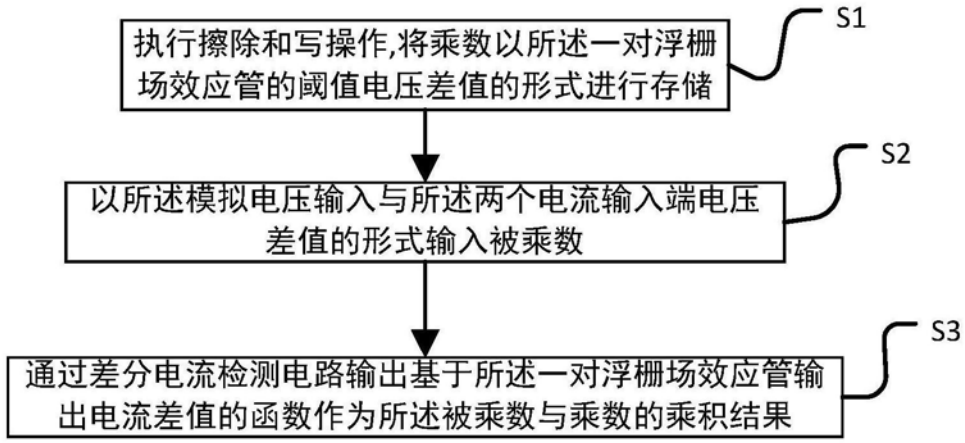

执行擦除和写操作,将乘数以所述一对浮栅场效应管的阈值电压差值的形式进行存储
S1
以所述模拟电压输入与所述两个电流输入端电压差值的形式输入被乘数
S2
通过差分电流检测电路输出基于所述一对浮栅场效应管输出电流差值的函数作为所述被乘数与乘数的乘积结果
S3

图10

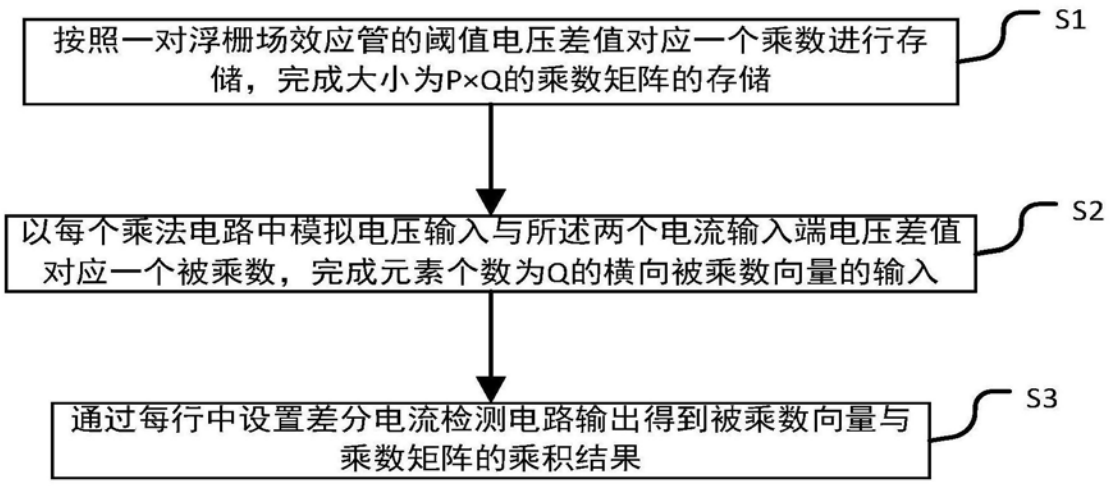

按照一对浮栅场效应管的阈值电压差值对应一个乘数进行存储，完成大小为P×Q的乘数矩阵的存储
S1
以每个乘法电路中模拟电压输入与所述两个电流输入端电压差值对应一个被乘数，完成元素个数为Q的横向被乘数向量的输入
S2
通过每行中设置差分电流检测电路输出得到被乘数向量与乘数矩阵的乘积结果
S3

图11

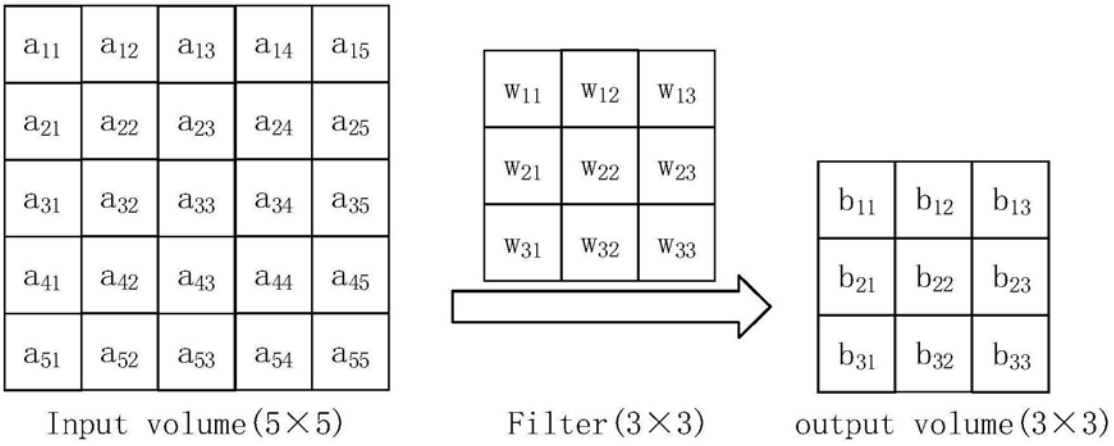

a11 a12 a13 a14 a15
a21 a22 a23 a24 a25
a31 a32 a33 a34 a35
a41 a42 a43 a44 a45
a51 a52 a53 a54 a55
Input volume(5×5)
w11 w12 w13
w21 w22 w23
w31 w32 w33
Filter(3×3)
b11 b12 b13
b21 b22 b23
b31 b32 b33
output volume(3×3)

图12

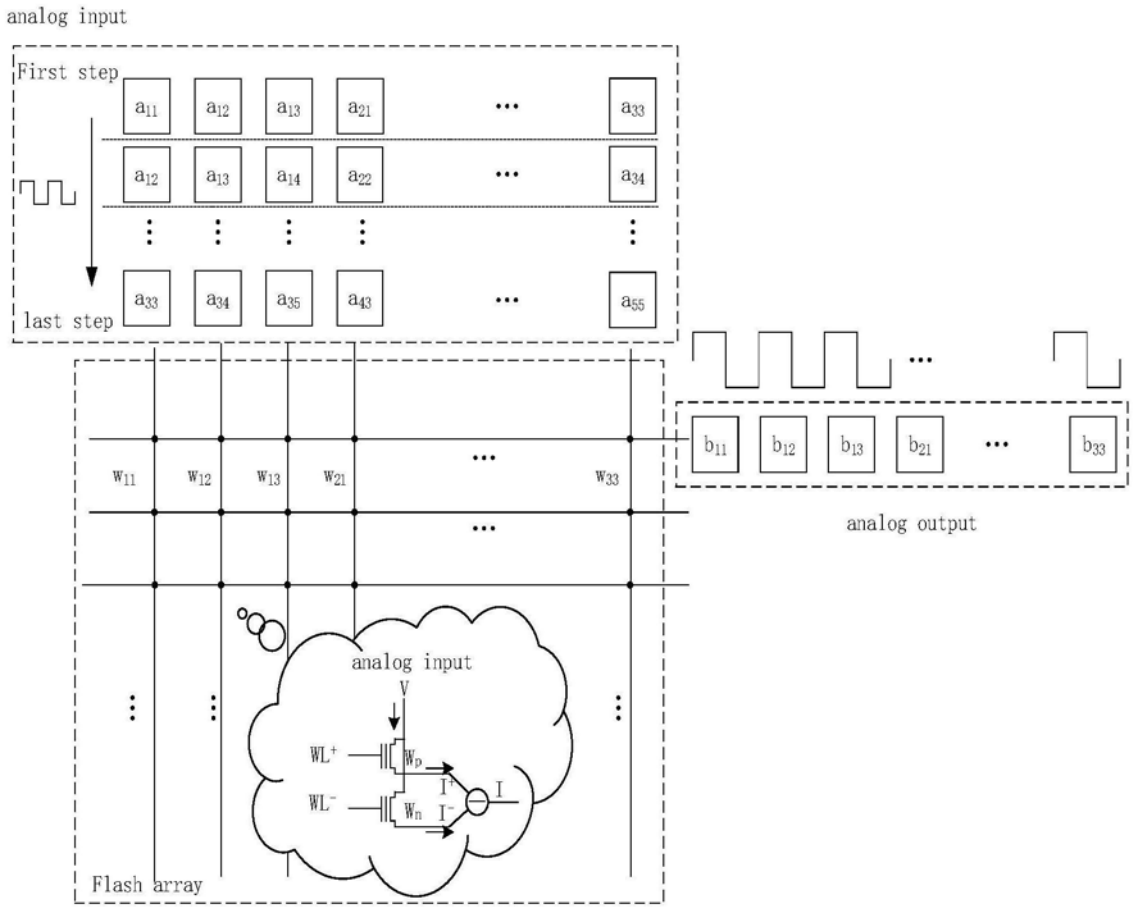

analog input
First step
last step
analog output
analog input
V
Flash array

图13

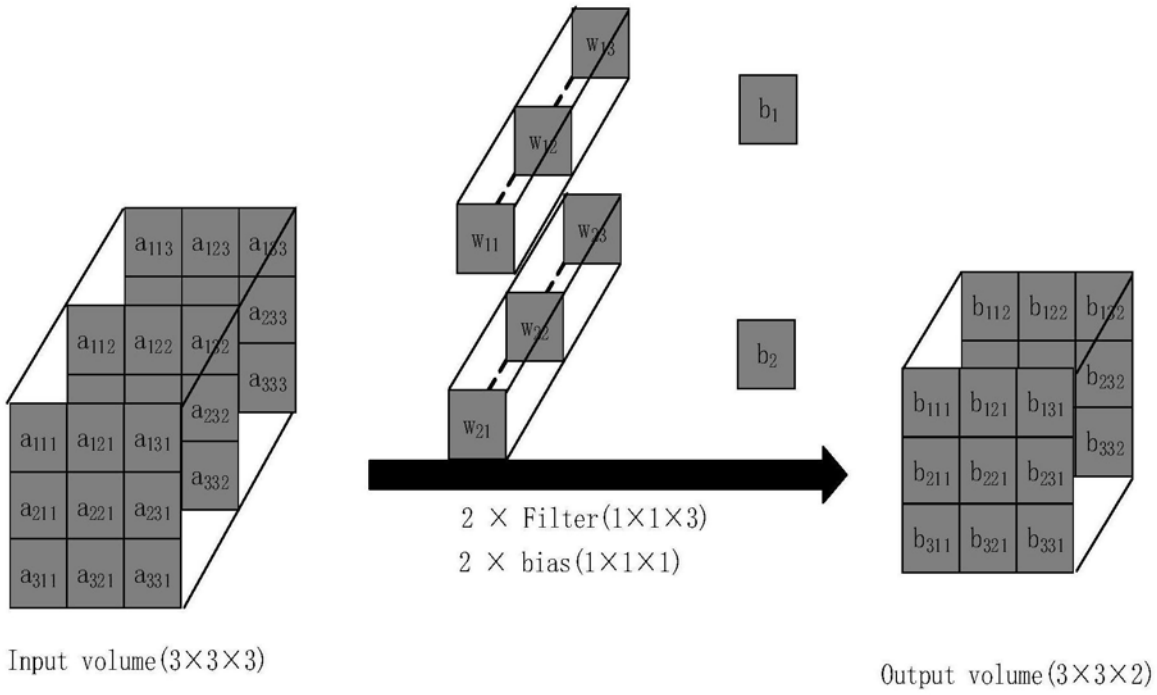

Input volume(3×3×3)
2 × Filter(1×1×3)
2 × bias(1×1×1)
Output volume(3×3×2)

图14

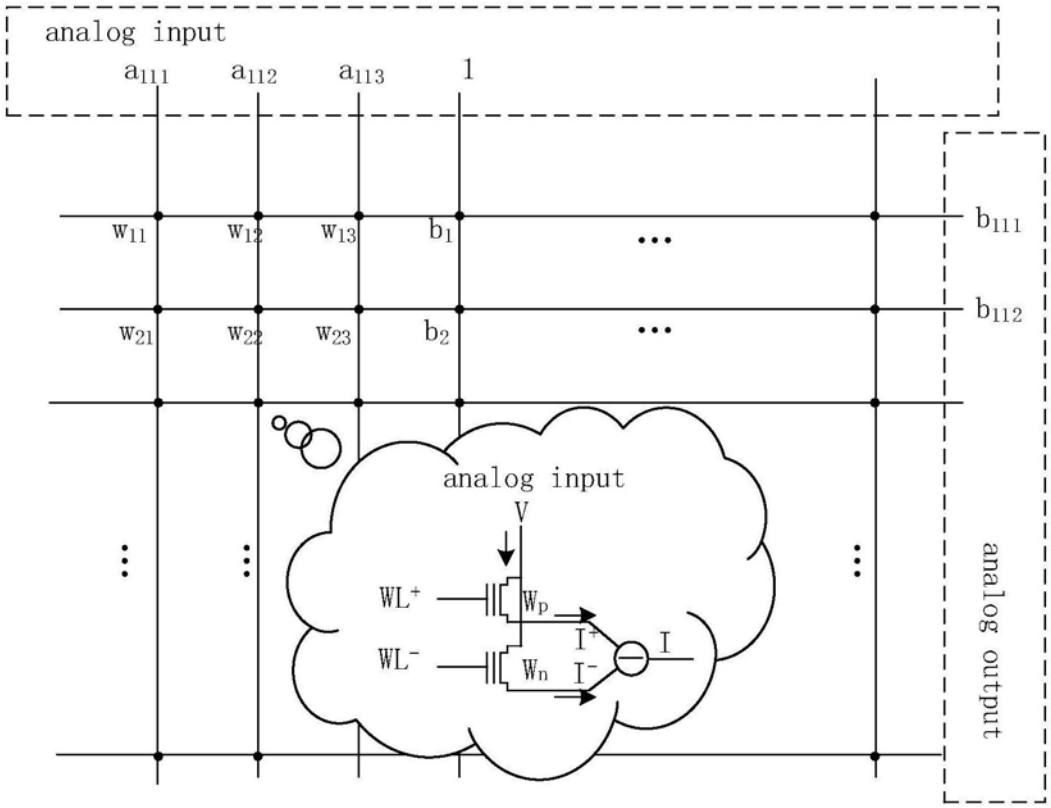

analog input
analog input
V
WL+
WL-
Wp
Wn
I+
I-
I
analog output

图15

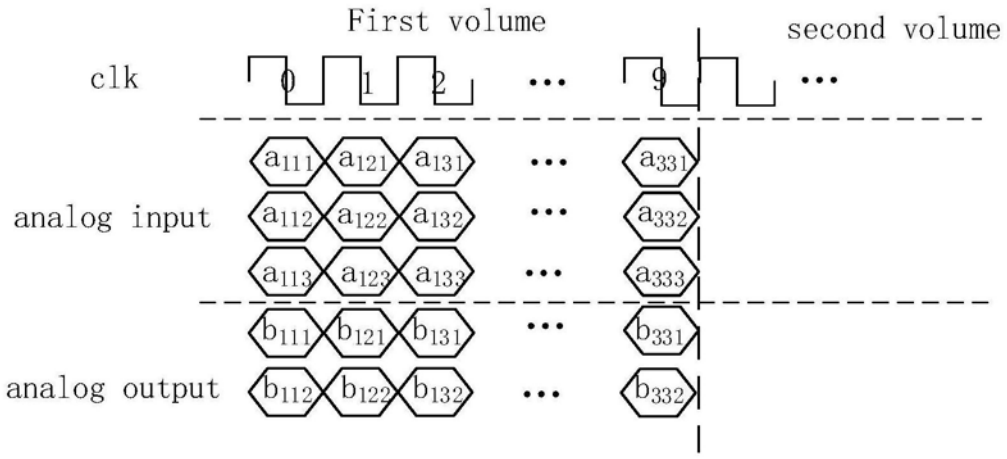

First volume
second volume
clk
0
1
2
9
analog input
analog output
a111 a121 a131 a331
a112 a122 a132 a332
a113 a123 a133 a333
b111 b121 b131 b331
b112 b122 b132 b332

图16

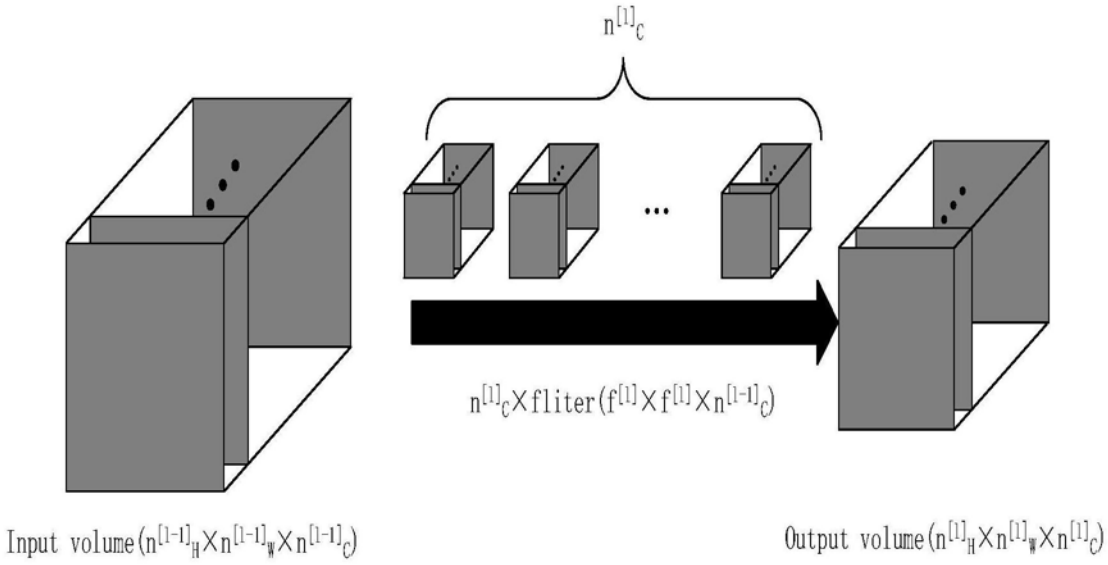

n[l]C
n[l]C×fliter(f[l]×f[l]×n[l-l]C)
Input volume(n[l-l]H×n[l-l]W×n[l-l]C)
Output volume(n[l]H×n[l]W×n[l]C)

图17

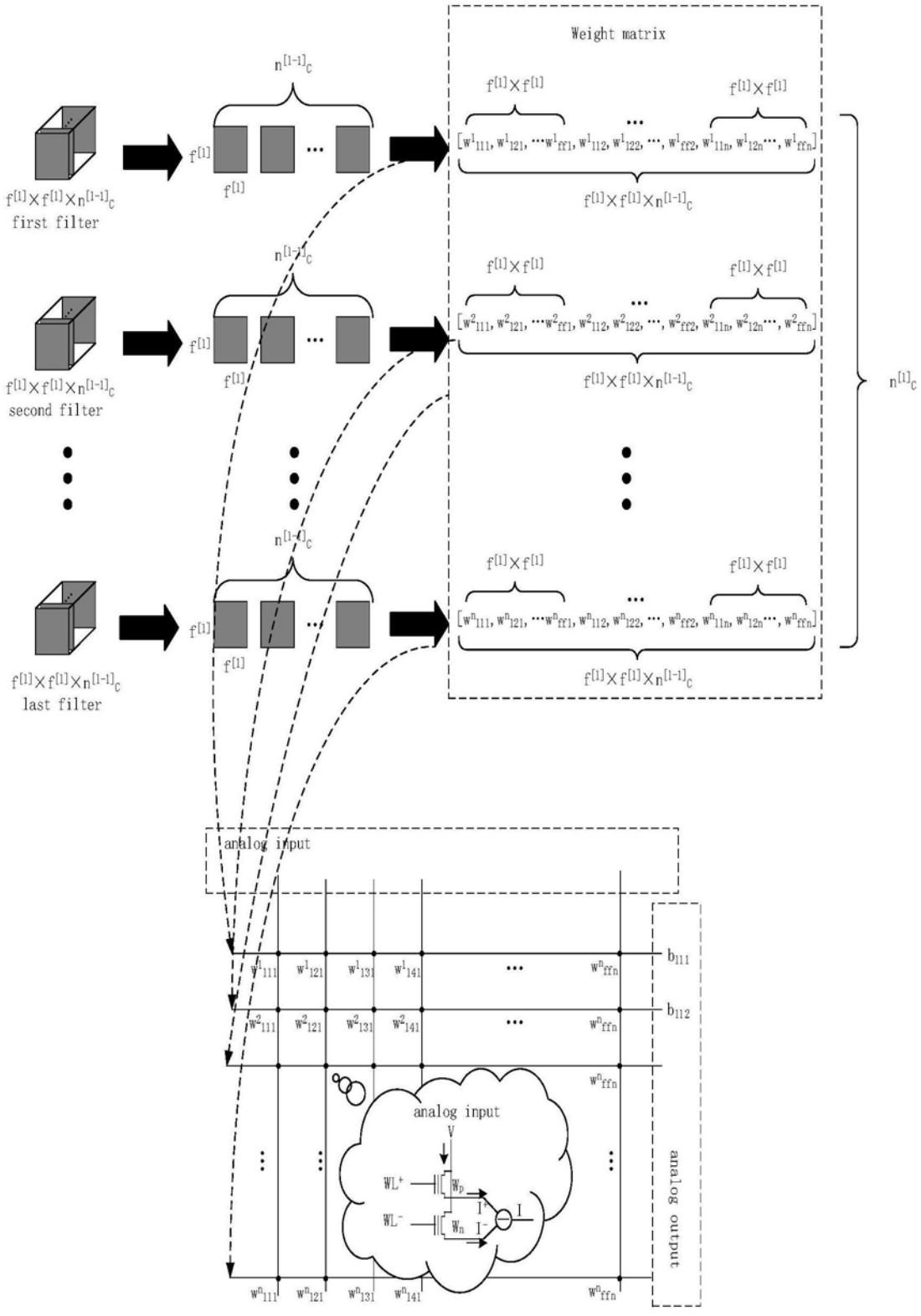

Weight matrix
first filter
second filter
last filter
analog input
analog output
WL+
WL-

图18

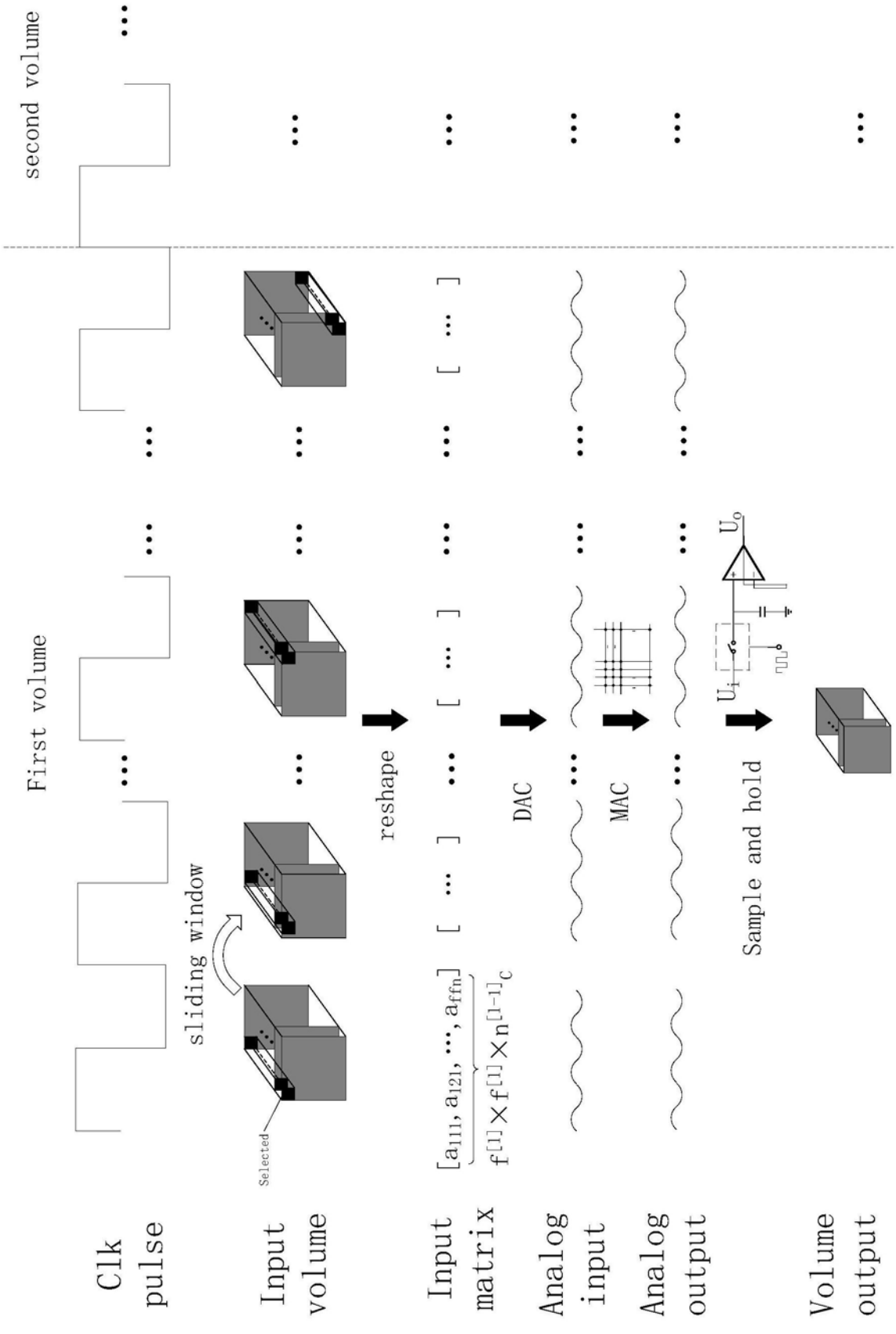

Clk pulse
Input volume
Input matrix
Analog input
Analog output
Volume output
First volume
second volume
sliding window
Selected
reshape
DAC
MAC
Sample and hold
$[a_{111}, a_{121}, \cdots, a_{ffn}]$
$f^{[l]} \times f^{[l]} \times n^{[l-1]}{}_C$
U_i
U_o

图19

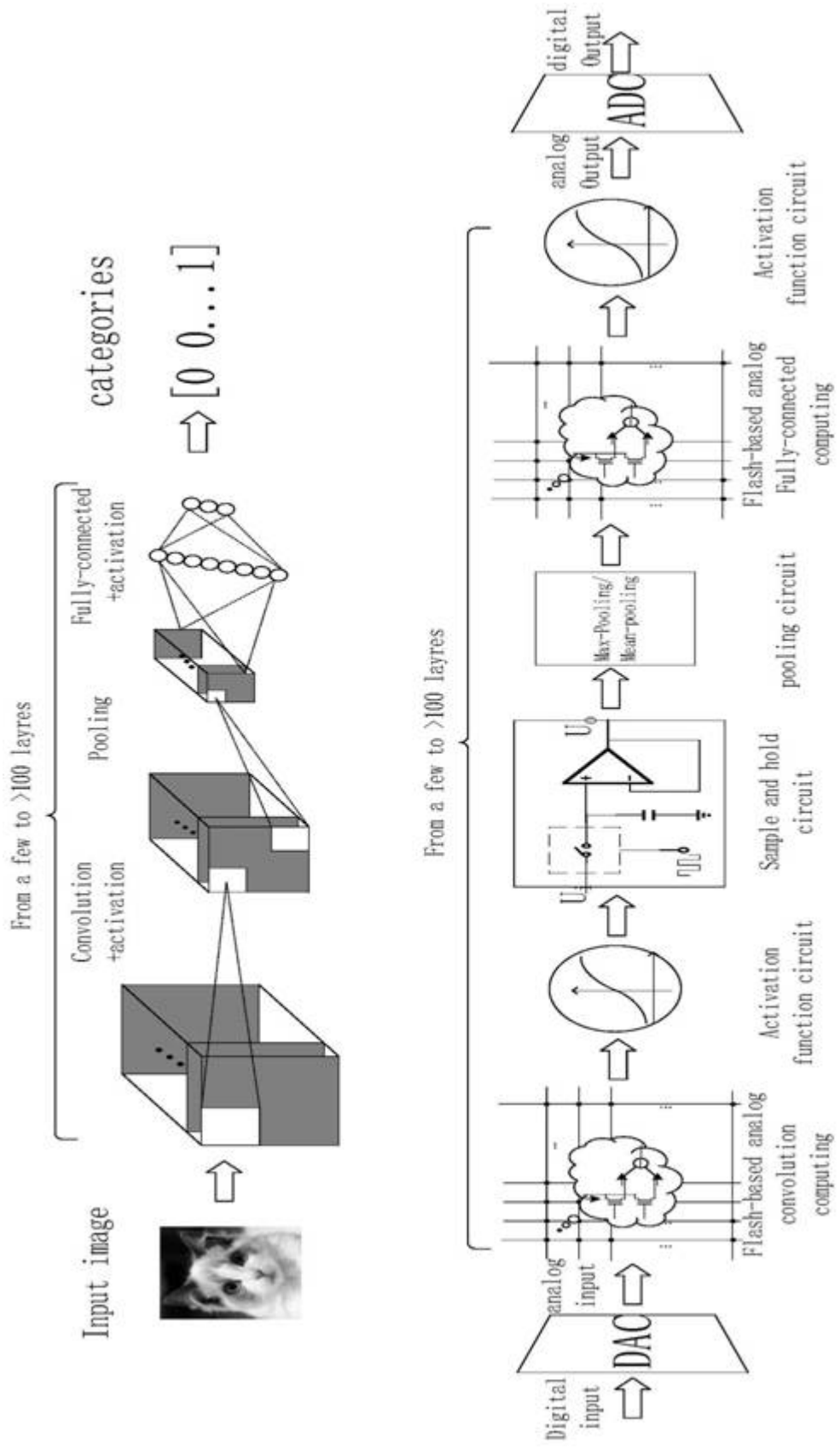

Input image
From a few to >100 layres
Convolution +activation
Pooling
Fully-connected +activation
categories
[0 0...1]
Digital input
DAC
analog input
Flash-based analog convolution computing
Activation function circuit
Sample and hold circuit
Max-Pooling/ Mean-pooling
pooling circuit
Flash-based analog Fully-connected computing
Activation function circuit
analog Output
ADC
digital Output

图20

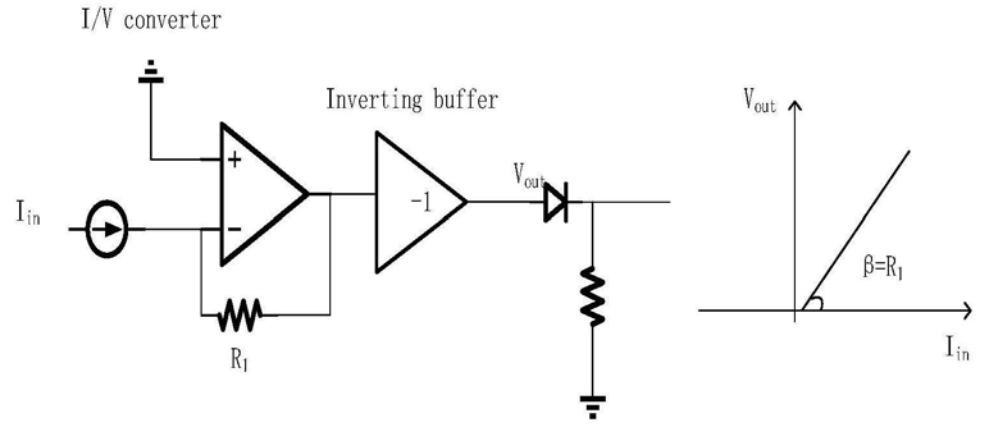

I/V converter
Inverting buffer
I_{in}
V_{out}
-1
R_1
V_{out}
β=R_1
I_{in}

图21

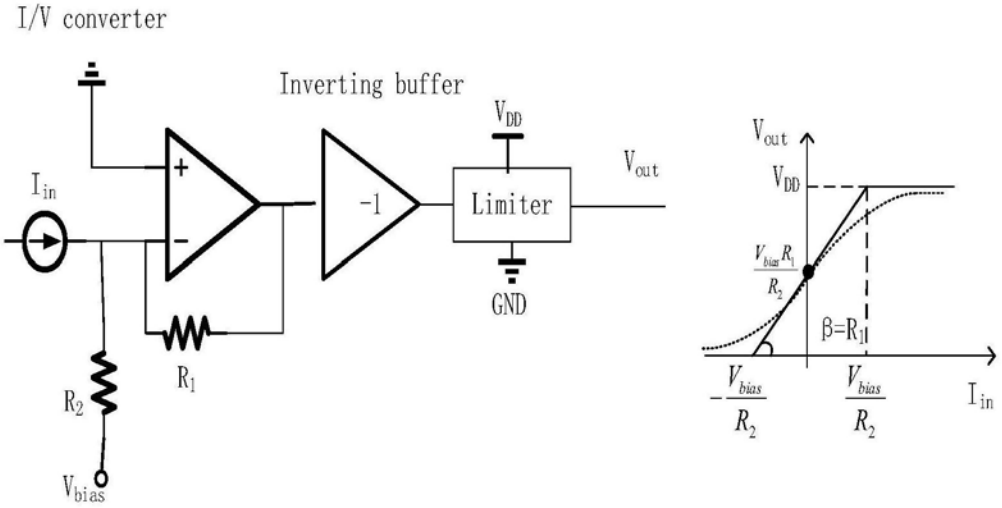

I/V converter
Inverting buffer
I_{in}
V_{DD}
V_{out}
-1
Limiter
GND
R_1
R_2
V_{bias}
V_{out}
V_{DD}
$\frac{V_{bias}R_1}{R_2}$
β=R_1
$-\frac{V_{bias}}{R_2}$
$\frac{V_{bias}}{R_2}$
I_{in}

图22

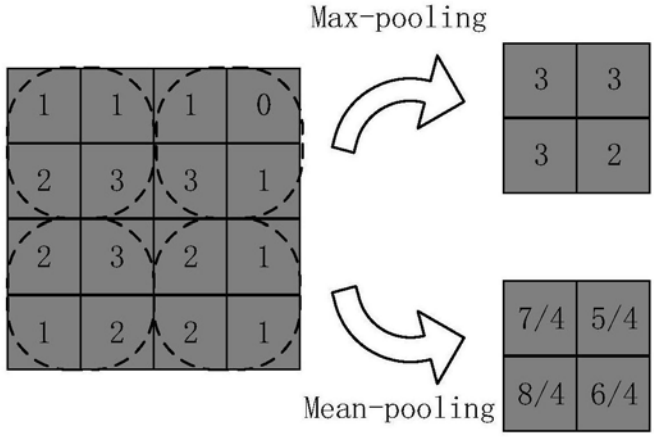

Max-pooling
1 1 1 0
2 3 3 1
2 3 2 1
1 2 2 1
3 3
3 2
7/4 5/4
8/4 6/4
Mean-pooling

图23

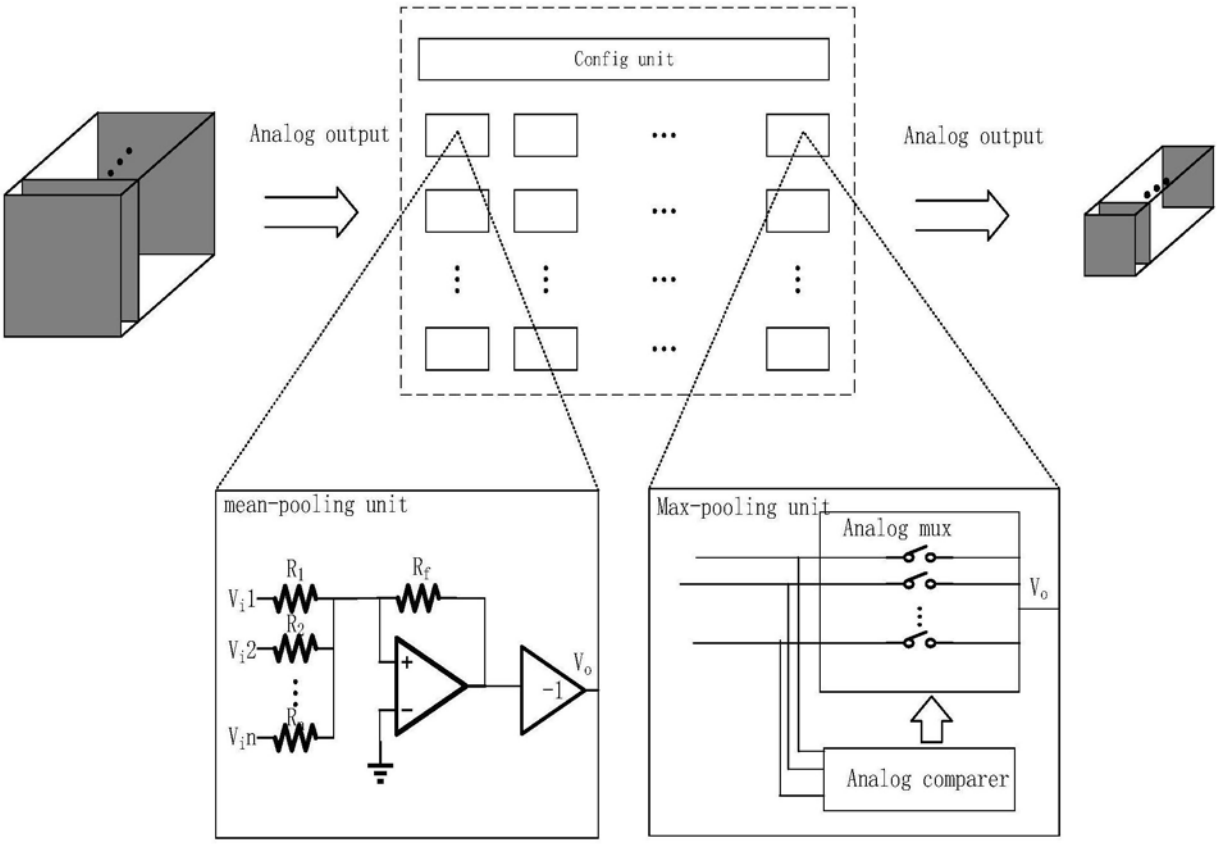

Config unit
Analog output
Analog output
mean-pooling unit
Max-pooling unit
Analog mux
Analog comparer
-1

图24

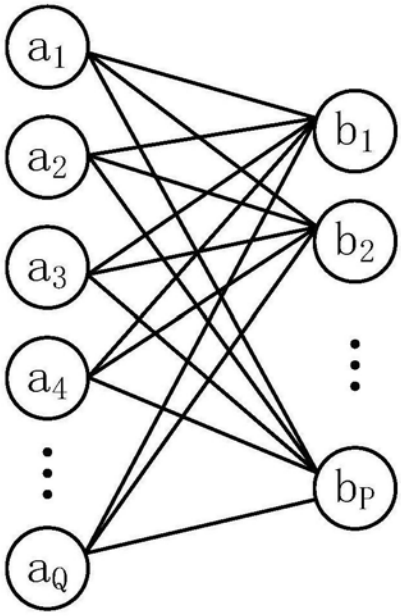

a1
a2
a3
a4
aQ
b1
b2
bP

图25

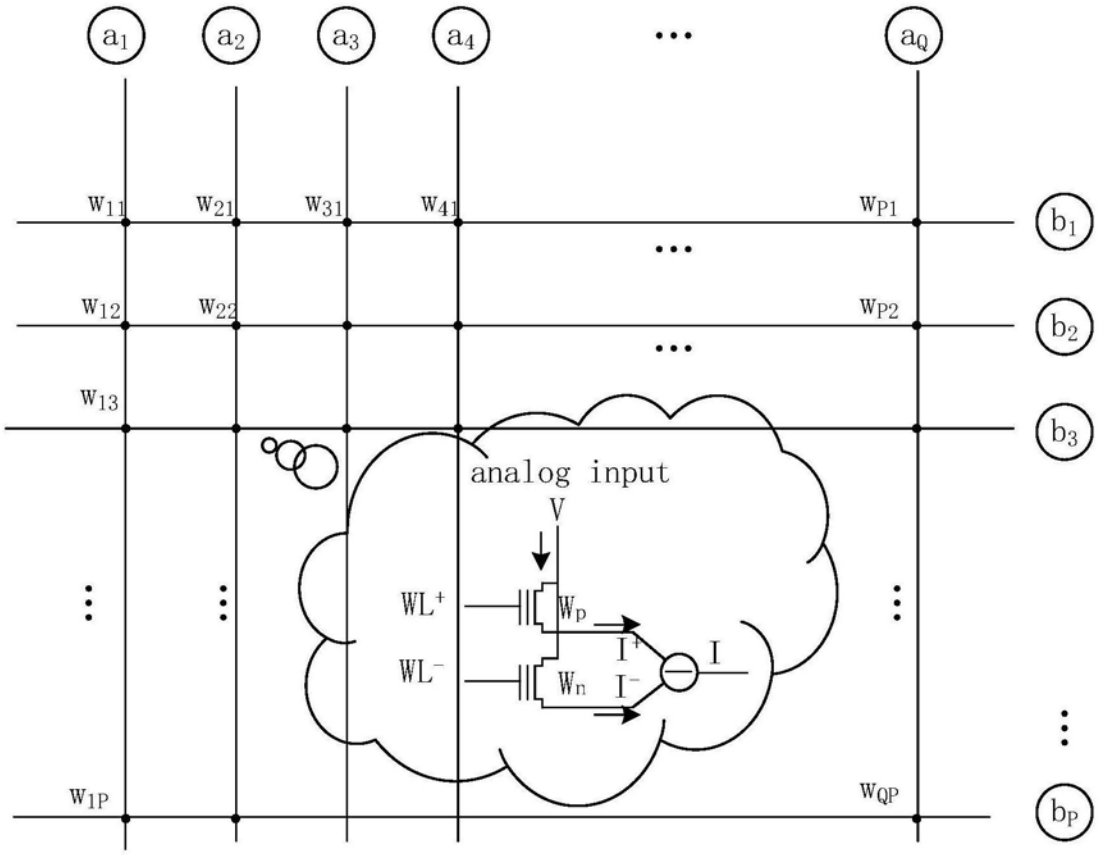

a1
a2
a3
a4
aQ
W11
W21
W31
W41
WP1
b1
W12
W22
WP2
b2
W13
b3
analog input
V
WL+
Wp
I+
I
WL−
Wn
I−
W1P
WQP
bP

图26