



(12)发明专利申请

(10)申请公布号 CN 111710356 A

(43)申请公布日 2020.09.25

(21)申请号 202010472550.9

G06N 3/08(2006.01)

(22)申请日 2020.05.29

(71)申请人 北京大学

地址 100871 北京市海淀区颐和园路5号

(72)发明人 黄鹏 项亚臣 康晋锋 刘晓彦

(74)专利代理机构 中科专利商标代理有限责任
公司 11021

代理人 李佳

(51)Int.Cl.

G11C 16/04(2006.01)

G11C 16/08(2006.01)

G11C 16/24(2006.01)

G11C 16/26(2006.01)

G06F 7/501(2006.01)

G06N 3/063(2006.01)

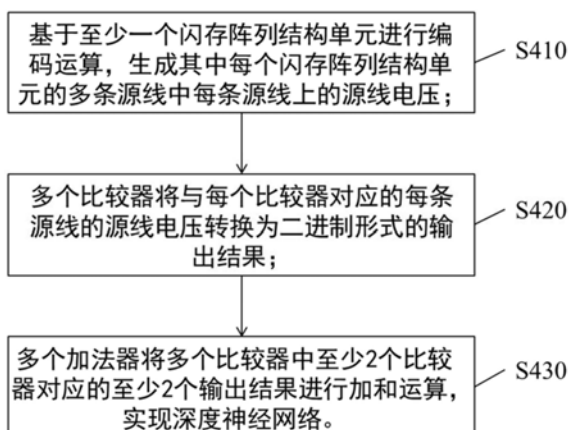
权利要求书2页 说明书11页 附图6页

(54)发明名称

编码型闪存装置和编码方法

(57)摘要

本发明公开了一种编码型闪存装置和编码方法,其中,编码型闪存装置包括:至少一个闪存阵列结构单元、多个比较器和多个加法器,至少一个闪存阵列结构单元中每个闪存阵列结构单元为3D NAND FLASH阵列结构单元,用于实现编码运算以生成闪存阵列结构单元的多条源线中每条源线上的源线电压;多个比较器中每个比较器与每条源线对应相连,用于将对应相连的每条源线的源线电压转换为二进制形式的输出结果;以及多个加法器中每个加法器与多个比较器中的至少2个比较器通过对应的每条源线相连,用于将至少2个比较器对应的至少两个输出结果进行加和运算。本发明的编码型闪存装置和编码方法可以实现高效且精确的全连接层或卷积层运算,从而实现深度神经网络。



1. 一种编码型闪存装置,其特征在于,包括:

至少一个闪存阵列结构单元,其中每个所述闪存阵列结构单元为3D NAND FLASH阵列结构单元,用于实现编码运算以生成所述闪存阵列结构单元的多条源线中每条源线上的源线电压;

多个比较器,其中每个比较器与所述每条源线对应相连,用于将对应相连的所述每条源线的源线电压转换为二进制形式的输出结果;以及

多个加法器,其中每个加法器与多个比较器中的至少2个比较器通过对应的所述每条源线相连,用于将所述至少2个比较器对应的至少两个输出结果进行加和运算,以实现深度神经网络。

2. 根据权利要求1所述的编码型闪存装置,其特征在于,每个所述闪存阵列结构单元包括:

多个运算单元,其中每个运算单元为多条字线中每条字线与所述每条源线在所述闪存阵列结构单元中交叉位置的晶体管,其中所述每条源线上对应设置一个所述运算单元;

多个冗余单元,为所述每个所述闪存阵列结构单元中的非运算单元的晶体管,用于在实现所述编码运算时处于开态。

3. 根据权利要求2所述的编码型闪存装置,其特征在于,每个所述闪存阵列结构单元还包括:

串选择线,与所述闪存阵列结构单元的位线端相连;

地选择线,与所述闪存阵列结构单元的源线端相连;

其中,所述串选择线与所述地选择线用于在实现所述编码运算时施加高电平。

4. 根据权利要求1所述编码型闪存装置,其特征在于,所述编码运算为全连接运算或卷积层运算,

所述多个加法器的数量与所述全连接运算或卷积层运算的加和运算的运算结果的数量一致。

5. 一种编码方法,基于权利要求1-4中任一项所述的编码型闪存装置实现,其特征在于,所述编码方法包括:

基于至少一个闪存阵列结构单元进行编码运算,生成其中每个闪存阵列结构单元的多条源线中每条源线上的源线电压;

多个比较器将与每个比较器对应的所述每条源线的源线电压转换为二进制形式的输出结果;以及

多个加法器将所述多个比较器中至少2个比较器对应的至少2个输出结果进行加和运算,实现深度神经网络。

6. 根据权利要求5所述的编码方法,其特征在于,所述编码运算为全连接运算或卷积层运算。

7. 根据权利要求6所述的编码方法,其特征在于,所述编码运算为全连接运算时,在至少一个闪存阵列结构单元进行编码运算生成其中每个闪存阵列结构单元的多条源线中每条源线上的源线电压之前,还包括:

将所述全连接层的权值矩阵向量中的多个非零位元素中的每个非零位元素预存入所述编码型闪存装置中多个运算单元中的对应运算单元中。

8. 根据权利要求6所述的编码方法,其特征在于,所述编码运算为卷积层运算时,在至少一个闪存阵列结构单元进行编码运算生成其中每个闪存阵列结构单元的多条源线中每条源线上的源线电压之前,还包括:

将所述卷积层的卷积矩阵中的多个卷积核元素中的每个卷积核元素预存入所述编码型闪存装置中多个运算单元中的对应运算单元中。

9. 根据权利要求7或8所述的编码方法,其特征在于,

在将所述全连接层的权值矩阵向量中的多个非零位元素中的每个非零位元素预存入所述编码型闪存装置中多个运算单元中的对应运算单元中之后,或在将所述卷积层的卷积矩阵中的多个卷积核元素中的每个卷积核元素预存入所述编码型闪存装置中多个运算单元中的对应运算单元中之后,还包括:

将对应的输入向量的输入元素对应输入多条字线中的每条字线中,以生成相应每条字线上的字线电压,使得所述编码型闪存装置的多个冗余单元处于开态。

10. 根据权利要求9所述的编码方法,其特征在于,在将对应的输入向量的输入元素对应输入多条字线中的每条字线中,以生成相应每条字线上的字线电压,使得所述编码型闪存装置的多个冗余单元处于开态之后,还包括:

对所述编码型闪存装置的串选择线和地选择线施加高电平,以生成其中每个闪存阵列结构单元的多条源线中每条源线上的源线电压。

11. 根据权利要求5所述的编码方法,其特征在于,所述编码运算为全连接运算时,所述基于至少一个闪存阵列结构单元进行编码运算,生成其中每个闪存阵列结构单元的多条源线中每条源线上的源线电压,包括:

基于多个闪存阵列结构单元同时进行对应输入向量和权值矩阵向量的编码运算,或者基于一个闪存阵列结构单元采用分时复用的方式进行对应输入向量和权值矩阵向量的编码运算;

其中,所述输入向量为多比特数据。

12. 根据权利要求5所述的编码方法,其特征在于,所述编码运算为卷积层运算时,所述编码方法还包括:

对所述多个加法器的加和运算输出的加和结果进行移位求和,实现深度神经网络;

其中,所述卷积层的卷积矩阵为多比特数据。

编码型闪存装置和编码方法

技术领域

[0001] 本发明涉及半导体器件与集成电路技术领域,特别涉及一种实现深度神经网络的编码型闪存装置和编码方法。

背景技术

[0002] 深度神经网络目前被广泛应用于图像处理和语音识别等领域并展现出了优异的性能。现有技术中的深度神经网络多依托于传统的冯诺依曼式架构进行运算,但受限于冯诺依曼式计算架构中存储单元和计算单元的分离,数据通过总线传输会导致延时增加、能耗增大,从而使得深度神经网络的数据处理能力和能效比的提高存在瓶颈。为此,现有技术进一步提出了基于3D NAND FLASH结构进行模拟运算的编码型闪存装置。但该编码型闪存装置,在实现深度神经网络的过程中,仍然存在三个主要问题如下:

[0003] 第一,当不同器件之间的阈值电压涨落较高时,计算准确率会下降;

[0004] 第二,模拟运算的结果需要经过复杂的模数转换电路,一定程度上降低了3D NAND FLASH存内运算所带来的能效比;

[0005] 第三,受限于3D NAND FLASH的操作方式,即同一时刻只能对单根字线WL上的数据进行运算,整个计算架构的效率并没有最大化、并行度也并未达到饱和。

发明内容

[0006] (一)要解决的技术问题

[0007] 为解决现有技术中,基于3D NAND FLASH结构进行模拟运算的编码型闪存装置存在计算准确率低、能效比降低以及计算架构效率没有最大化且并行度未达到饱和等技术问题,本发明提供了一种编码型闪存装置和编码方法。

[0008] (二)技术方案

[0009] 本发明的一个方面公开了一种编码型闪存装置,其中,包括:至少一个闪存阵列结构单元、多个比较器和多个加法器,至少一个闪存阵列结构单元中每个闪存阵列结构单元为3D NAND FLASH阵列结构单元,用于实现编码运算以生成闪存阵列结构单元的多条源线中每条源线上的源线电压;多个比较器中每个比较器与每条源线对应相连,用于将对应相连的每条源线的源线电压转换为二进制形式的输出结果;以及多个加法器中每个加法器与多个比较器中的至少2个比较器通过对应的每条源线相连,用于将至少2个比较器对应的至少两个输出结果进行加和运算,以实现深度神经网络。

[0010] 根据本发明的实施例,其中,每个闪存阵列结构单元包括:多个运算单元和多个冗余单元;多个运算单元中每个运算单元为多条字线中每条字线与每条源线在闪存阵列结构单元中交叉位置的晶体管,其中每条源线上对应设置一个运算单元;多个冗余单元为每个闪存阵列结构单元中的非运算单元的晶体管,用于在实现编码运算时处于开态。

[0011] 根据本发明的实施例,其中,每个闪存阵列结构单元还包括:串选择线和地选择线,串选择线与闪存阵列结构单元的位线端相连;地选择线与闪存阵列结构单元的源线端

相连;其中,串选择线与地选择线用于在实现编码运算时施加高电平。

[0012] 根据本发明的实施例,其中,编码运算为全连接运算或卷积层运算,多个加法器的数量与全连接运算或卷积层运算的加和运算的运算结果的数量一致。

[0013] 本发明的另一个方面公开了一种编码方法,基于上述的编码型闪存装置实现,其中,编码方法包括:基于至少一个闪存阵列结构单元进行编码运算,生成其中每个闪存阵列结构单元的多条源线中每条源线上的源线电压;多个比较器将与每个比较器对应的每条源线的源线电压转换为二进制形式的输出结果;以及多个加法器将多个比较器中至少2个比较器对应的至少2个输出结果进行加和运算,实现深度神经网络。

[0014] 根据本发明的实施例,其中,编码运算为全连接运算或卷积层运算。

[0015] 根据本发明的实施例,其中,编码运算为全连接运算时,在至少一个闪存阵列结构单元进行编码运算生成其中每个闪存阵列结构单元的多条源线中每条源线上的源线电压之前,编码方法还包括:将全连接层的权值矩阵向量中的多个非零位元素中的每个非零位元素预存入编码型闪存装置中多个运算单元中的对应运算单元中。

[0016] 根据本发明的实施例,其中,编码运算为卷积层运算时,在至少一个闪存阵列结构单元进行编码运算生成其中每个闪存阵列结构单元的多条源线中每条源线上的源线电压之前,编码方法还包括:将卷积层的卷积矩阵中的多个卷积核元素中的每个卷积核元素预存入编码型闪存装置中多个运算单元中的对应运算单元中。

[0017] 根据本发明的实施例,其中,在将全连接层的权值矩阵向量中的多个非零位元素中的每个非零位元素预存入编码型闪存装置中多个运算单元中的对应运算单元中之后,或在将卷积层的卷积矩阵中的多个卷积核元素中的每个卷积核元素预存入编码型闪存装置中多个运算单元中的对应运算单元中之后,编码方法还包括:将对应的输入向量的输入元素对应输入多条字线中的每条字线中,以生成相应每条字线上的字线电压,使得编码型闪存装置的多个冗余单元处于开态。

[0018] 根据本发明的实施例,其中,在将对应的输入向量的输入元素对应输入多条字线中的每条字线中,以生成相应每条字线上的字线电压,使得编码型闪存装置的多个冗余单元处于开态之后,编码方法还包括:对编码型闪存装置的串选择线和地选择线施加高电平,以生成其中每个闪存阵列结构单元的多条源线中每条源线上的源线电压。

[0019] 根据本发明的实施例,其中,编码运算为全连接运算时,基于至少一个闪存阵列结构单元进行编码运算,生成其中每个闪存阵列结构单元的多条源线中每条源线上的源线电压,包括:基于多个闪存阵列结构单元同时进行对应输入向量和权值矩阵向量的编码运算,或者基于一个闪存阵列结构单元采用分时复用的方式进行对应输入向量和权值矩阵向量的编码运算;其中,输入向量为多比特数据。

[0020] 根据本发明的实施例,其中,编码运算为卷积层运算时,编码方法还包括:对多个加法器的加和运算输出的加和结果进行移位求和,实现深度神经网络;其中,卷积层的卷积矩阵为多比特数据。

[0021] (三)有益效果

[0022] 本发明公开了一种编码型闪存装置和编码方法,其中,编码型闪存装置包括:至少一个闪存阵列结构单元、多个比较器和多个加法器,至少一个闪存阵列结构单元中每个闪存阵列结构单元为3D NAND FLASH阵列结构单元,用于实现编码运算以生成闪存阵列结构

单元的多条源线中每条源线上的源线电压；多个比较器中每个比较器与每条源线对应相连，用于将对应相连的每条源线的源线电压转换为二进制形式的输出结果；以及多个加法器中每个加法器与多个比较器中的至少2个比较器通过对应的每条源线相连，用于将至少2个比较器对应的至少两个输出结果进行加和运算。本发明的编码型闪存装置和编码方法可以实现高效且精确的全连接层或卷积层运算，从而实现深度神经网络。

附图说明

[0023] 图1是根据本发明一实施例的对应单个闪存阵列结构单元的编码型闪存装置的组成示意图；

[0024] 图2是根据本发明实施例的编码型闪存装置的编码方法的流程示意图；

[0025] 图3根据本发明又一实施例的应用于全连接运算的编码型闪存装置的组成示意图；

[0026] 图4根据本发明又一实施例的应用于全连接运算的编码型闪存装置的编码方法的部分流程示意图；

[0027] 图5根据本发明另一实施例的应用于卷积层运算的编码型闪存装置的组成示意图；

[0028] 图6根据本发明另一实施例的应用于卷积层运算的编码型闪存装置的编码方法的部分流程示意图。

具体实施方式

[0029] 为使本发明的目的、技术方案和优点更加清楚明白，以下结合具体实施例，并参照附图，对本发明进一步详细说明。

[0030] 为解决现有技术中，基于3D NAND FLASH结构进行模拟运算的编码型闪存装置存在计算准确率低、能效比降低以及计算架构效率没有最大化且并行度未达到饱和等技术问题，本发明提供了一种编码型闪存装置和编码方法。

[0031] 本发明的编码型闪存装置和编码方法能够基于3D NAND FLASH结构实现并行存内计算的编码运算基于如下技术原理：在同一根字线WL上，当所有存储单元（即晶体管）均处于开态时，对应的源线SL才能读出源线电流。基于此原理，为了实现基于3D NAND FLASH结构的并行存内计算，区别于传统的串行操作字线WL实现3D NAND FLASH存内计算的编码方法，如图1所示，基于本发明的编码型闪存装置的编码方法技术构思如下：

[0032] 在3D NAND FLASH结构上每条源线SL上预设定仅有一个存储单元负责数据运算（即运算单元），

[0033] 该运算单元中可以存储的数值为0或1，其中分别对应于运算单元上的阈值电压为 $V_{th}=V_{th_Low}$ （即运算单元存储数值为0）或 $V_{th}=V_{th_High}$ （即运算单元存储数值为1）；

[0034] 在3D NAND FLASH结构上的其余存储单元（非运算单元）设置为冗余单元，冗余单元上存储的数值均为0，则对应冗余单元上的阈值电压为 $V_{th}=V_{th_Low}$ ；

[0035] 运算单元上输入数值0或1，则对应于运算单元所在的字线WL上施加的字线电压 $V_g=V_{g_High}$ 或 $V_g=V_{g_Low}$ ；

[0036] 可见，无论运算单元上输入数值为0或1，其他的冗余单元均处于开态，因此，源线

SL的源线电流将对应出现如下运算结果：

[0037] 当源线SL读出的源线电流为0时，表示运算单元处于截止态，即 $V_{th}=V_{th_High}(1)$ ， $V_g=V_{g_Low}(1)$ ，代表运算结果为1；

[0038] 当SL读出的源线电流为1时，表示运算单元处于开态，即存在以下三种情况之一： $V_{th}=V_{th_High}(1)$ ， $V_g=V_{g_High}(0)$ ； $V_{th}=V_{th_Low}(0)$ ， $V_g=V_{g_High}(0)$ ； $V_{th}=V_{th_Low}(0)$ ， $V_g=V_{g_Low}(1)$ ，代表运算结果为0。

[0039] 因此，通过采用本发明的编码型闪存装置，能够在3D NAND FLASH结构内实现并行存内计算的编码方法。具体地，本发明的编码型闪存装置通过加法器辅助的基于3D NAND FLASH的深度神经网络，能够利用存算一体技术并行完成各计算层的功能，并得到准确的运算结果。

[0040] 本发明的一个方面公开了一种编码型闪存装置，如图1、图3和图5所示，其中，编码型闪存装置包括：至少一个闪存阵列结构单元、多个比较器和多个加法器，用于实现上述的基于3D NAND FLASH的深度神经网络。

[0041] 编码型闪存装置包括：至少一个闪存阵列结构单元100，如图3和图4中彼此相互并联连接的闪存阵列结构单元100-1至闪存阵列结构单元110-N，共N个闪存阵列结构单元， $N \geq 1$ 。N个闪存阵列结构单元中每个闪存阵列结构单元100为3D NAND FLASH阵列结构单元，该3D NAND FLASH阵列结构单元同时具有多个存储单元、多条位线BL、多条字线WL和多条源线SL。

[0042] 如图1所示，该多个存储单元的数量可以是 $M \times M$ 个，其中M个存储单元按照一定的排列规则相互串联形成M个存储单元串，每个存储单元串的一端对应连接一条位线BL，另一端对应连接一条源线SL，每条字线WL与存储单元串相互垂直，连接多个存储单元串上对应的存储单元，因此，位线BL、源线SL的数量为M条，字线WL的数量对应存储单元串上串联的存储单元数量也为M条，每个闪存阵列结构单元100用于实现编码运算以生成闪存阵列结构单元的多条源线中每条源线上的源线电压。

[0043] 具体地，在对应每条源线SL上预设定的运算单元存储数值0或1，其他的冗余单元存储数值为0，此时冗余单元将处于开态，对应于运算单元所在的字线WL上施加的字线电压为 $V_g=V_{g_High}$ 或 $V_g=V_{g_Low}$ ，此时即可生成与运算单元所在的源线SL相应的源线电压。

[0044] 多个比较器中每个比较器与每条源线对应相连，用于将对应相连的每条源线的源线电压转换为二进制形式的输出结果，本发明的比较器可以为集成电路领域的标准器件，为将模拟电压信号与基准电压相比较的电路，可以在两路输入为模拟信号，实现输出为二进制数值0或1，且在输入电压的发生数值波动时，输出值保持恒定。在本发明的另一实施例中，上述比较器可以替换为采样电路。

[0045] 多个加法器中每个加法器与多个比较器中的至少2个比较器通过对应的每条源线相连，用于将至少2个比较器对应的至少两个输出结果进行加和运算，以实现深度神经网络。加法器为计算机技术领域中的逻辑器件，用于执行数字的加和运算。该深度神经网络中包括多个卷积层数据或全连接层数据的处理。

[0046] 因此，本发明的编码型闪存装置通过比较器输出二进制形式的输出结果，不仅能够不同器件之间的阈值电压涨落较高时，维持较高的计算准确率，而且还能够避免类似传统输出结果数值的复杂性，避免传统模拟运算所需要的复杂模数转换电路，提高了3D

NAND FLASH结构存内运算所带来的能效比,最后借助于加法器的辅助,实现了最大化整个计算架构的效率和并行度,可在一个时钟周期之内完成并行的编码运算。

[0047] 根据本发明的实施例,其中,如图1、图3和图5所示,每个闪存阵列结构单元包括:多个运算单元C和多个冗余单元;多个运算单元C中每个运算单元C为多条字线中每条字线与每条源线在闪存阵列结构单元中交叉位置的晶体管,其中每条源线上对应设置一个运算单元C;运算单元C为在3D NAND FLASH结构上每条源线SL上的存储单元串里预设的一个存储单元,并且该源线SL存储单元串里仅由这个存储单元负责数据运算。该运算单元C中可以存储的数值为0或1,其中分别对应于运算单元C上的阈值电压为 $V_{th}=V_{th_Low}$ (即运算单元C存储数值为0)或 $V_{th}=V_{th_High}$ (即运算单元C存储数值为1)。运算单元用于全连接层运算或卷积层运算,在全连接层运算时,全连接层的权值矩阵向量中的多个非零位元素中的每个非零位元素对应存储到相应运算单元C中;在进行卷积层运算时,将卷积层的卷积矩阵中的多个卷积核元素中的每个卷积核元素对应存储到相应运算单元C中。

[0048] 多个冗余单元为每个闪存阵列结构单元100中的非运算单元C的晶体管,用于在实现编码运算时处于开态。冗余单元为对应于上述每条源线上存储单元串中非运算单元C的存储单元,相当于处于非每条字线与每条源线交叉位置的晶体管即冗余单元。当本发明的编码型闪存装置进行编码预算时,无论运算单元C上输入数值为0或1,冗余单元均处于开态。

[0049] 根据本发明的实施例,如图1、图3和图5所示,其中,每个闪存阵列结构单元100还包括:串选择线SSL(serial select line)和地选择线GSL(ground select line),串选择线SSL与闪存阵列结构单元100的位线端相连,具体地,串选择线SSL上设置有多个选择管,每个选择管的一端连接位线BL,另一端连接存储单元串的位端,串选择线SSL将多个选择管沿与源线SL相互垂直的方向串接。

[0050] 地选择线GSL与闪存阵列结构单元100的源线端相连;具体地,地选择线GSL上设置有多个选择管,每个选择管的一端连接源线SL,另一端连接存储单元串的源端,地选择线GSL将多个选择管沿与源线SL相互垂直的方向串接。

[0051] 其中,串选择线SSL与地选择线GSL用于在实现编码运算时施加高电平。

[0052] 根据本发明的实施例,其中,编码运算为全连接运算或卷积层运算,多个加法器的数量与全连接运算或卷积层运算的加和运算的运算结果的数量一致。具体地,每个闪存阵列结构单元100上加法器的数量1与该闪存阵列结构单元100上的全连接运算或卷积层运算的加和运算的运算结果的数量一致,本发明的编码型闪存装置的加法器数量L则为 $L=1 \times N$,N为该编码型闪存装置中并联连接的闪存阵列结构单元100的数量。

[0053] 对于全连接层运算,如图3所示,输入向量 $X(1 \times M)$ 与权值矩阵向量 $K(M \times N)$ 相乘后得到输出向量 $Y(1 \times N)$,数学表达如下公式(1):

$$[0054] \quad Y_i = X_1 \cdot K_{1,i} + X_2 \cdot K_{2,i} + \dots + X_M \cdot K_{M,i}$$

[0055] 其中 $1 \leq i \leq N$ 。

[0056] 因此,全连接运算的输出结果Y可以具有N个,则对应全连接运算需要的加法器的数量等于N。

[0057] 对于卷积层运算,如图5所示,多个局部区域的全连接层,输入向量的矩阵 $X(M \times N)$ 的各部分分别与卷积矩阵 $K(k \times k)$ 进行矩阵向量乘法运算,完成卷积操作并得到输出向量

的矩阵Y ($m \times n$) ,数学表达如下公式 (2) :

$$[0058] \quad Y_{i,j} = X_{i,j} \cdot K_{k,k} + X_{i,j+1} \cdot K_{k,k-1} + X_{i+1,j} \cdot K_{k-1,k} + \dots + X_{i+k-1,j+k-1} \cdot K_{1,1}$$

[0059] 其中 $1 \leq i \leq M-k+1$ (m) , $1 \leq j \leq N-k+1$ (n) 。

[0060] 因此,卷积层运算的输出结果Y可以具有 $(M-k+1) \times (N-k+1)$ 个,其构成一个完整的输出矩阵,则对应卷积层运算所需要的加法器的数量等于 $(M-k+1) \times (N-k+1)$ 。

[0061] 无论是卷积层运算亦或是全连接运算,都相当于向量矩阵乘法运算。相比之下,卷积层比全连接层增加了卷积核的移位操作。在利用3D NAND FLASH结构的闪存阵列结构单元100实现两种类型的向量矩阵乘法运算时,主要差异在于卷积层的位线BL输入会在每个周期后更新为输入图像待卷积的下一个局部区域。

[0062] 以上已经结合附图1、图3和图5对本发明的编码型闪存装置的实施例进行了详细描述。

[0063] 现具体阐述将3D NAND FLASH应用于向量矩阵乘法运算、实现并行化数据处理的过程。

[0064] 本发明的另一个方面公开了一种编码方法,如图2所示,基于上述的编码型闪存装置实现,其中,该编码方法包括:

[0065] 步骤S410:基于至少一个闪存阵列结构单元进行编码运算,生成其中每个闪存阵列结构单元的多条源线中每条源线上的源线电压;具体地,在对应每条源线SL上预设定的运算单元存储数值0或1,其他的冗余单元存储数值为0,此时冗余单元将处于开态,对应于运算单元所在的字线WL上施加的字线电压为 $V_g = V_{g_High}$ 或 $V_g = V_{g_Low}$,此时即可生成与运算单元所在的源线SL相应的源线电压。

[0066] 步骤S420:多个比较器将与每个比较器对应的每条源线的源线电压转换为二进制形式的输出结果;本发明的比较器可以在两路输入为模拟信号,实现输出为二进制数值0或1,且在输入电压的发生数值波动时,输出值保持恒定。

[0067] 步骤S430:多个加法器将多个比较器中至少2个比较器对应的至少2个输出结果进行加和运算,实现深度神经网络。加法器为计算机技术领域中的逻辑器件,用于执行数字的加和运算。该深度神经网络中包括多个卷积层数据或全连接层数据的处理。

[0068] 基于本发明的编码型闪存装置的编码方法,能够在3D NAND FLASH结构内实现并行存内计算的编码方法。具体地,本发明的编码型闪存装置通过加法器辅助的基于3D NAND FLASH的深度神经网络能够利用存算一体技术并行完成各计算层的功能并得到准确的运算结果。

[0069] 因此,本发明的编码方法通过比较器输出二进制形式的输出结果,不仅能够不同器件之间的阈值电压涨落较高时,维持较高的计算准确率,而且还能够避免类似传统输出结果数值的复杂性,避免传统模拟运算所需要的复杂模数转换电路,提高了3D NAND FLASH结构存内运算所带来的能效比,最后借助于加法器的辅助,实现了最大化整个计算架构的效率和并行度,可在一个时钟周期之内完成并行的编码运算。

[0070] 以下对本发明具有单个闪存阵列结构单元的编码型闪存装置的编码方法作进一步的解释。

[0071] 根据本发明的一实施例,如图1所示,对于本发明的上述编码型闪存装置,其中某一闪存阵列结构单元100在编码运算开始之前,将权值矩阵向量K ($M \times N$) 的各个元素 ($K_1 \sim$

K_M) 中的非零位元素分别存入 3D NAND FLASH 结构单元的 M 条源线 SL 上的对应的运算单元 C 上, 即非零位元素 K_1 存入字线 WL1 和源线 SL1 的交叉位置, 非零位元素 K_2 存入字线 WL2 和源线 SL2 的交叉位置……依此类推, 非零位元素 K_M 存入字线 WL_M 和源线 SL_M 的交叉位置。

[0072] 如前述, 在运算单元 C 中存储的数值 0 或 1, 对应于的运算单元的阈值电压为 $V_{th} = V_{th_Low}$ 或 V_{th_High} , 每条源线 SL 上仅有一个器件为运算单元 C, 其余单元设置为冗余单元, 每个冗余单元存储的数值均为 0, 即对应冗余单元上的阈值电压 $V_{th} = V_{th_Low}$, 使得冗余单元保持开态。

[0073] 输入向量 $(1 \times M)$ 的各个元素 $(X_1 \sim X_M)$ 分别转化为对应的电压值并施加至字线 WL1 ~ 字线 WL-M 上, 对应电压值对应于字线 WL 上输入 0 或 1, 使得字线 WL 上施加的字线电压为 $V_g = V_{g_High}$ 或 V_{g_Low} 。借此, 可以保证冗余单元对应的字线 WL 的输入数据不会影响冗余单元的状态, 即冗余单元一直处于开态。

[0074] 最后, 在该闪存阵列结构单元 100 的串选择线 SSL 和地选择线 GSL 同时加高电平, 源线 SL 的输出此时仅取决于对应的运算单元 C 的状态, 具体而言: 当运算单元 C 中存储的权值和通过字线 WL 输入的输入数据均为 1 时, 即该运算单元 C 的阈值电压为 $V_{th} = V_{th_High}$, 其所在字线 WL 的字线电压 $V_g = V_{g_Low}$, 此时运算单元 C 处于截止态, 对应的源线 SL 无源线电流, 代表运算结果为 1; 其余情况: $V_{th} = V_{th_High}$ (1), $V_g = V_{g_High}$ (0); $V_{th} = V_{th_Low}$ (0), $V_g = V_{g_High}$ (0); $V_{th} = V_{th_Low}$ (0), $V_g = V_{g_Low}$ (1), 运算单元均处于开态, 对应的源线 SL 有源线电流, 代表运算结果为 0。

[0075] 通过源线 SL 末端与源线相连的比较器 200, 可以将源线电流的有无读出并转换为计算结果 0 或 1, 即作为比较器 200 二进制形式的输出结果; 不同源线 SL 的运算单元 C 得到的输出结果通过加法器进行加和运算, 即可得到输入数据和权值数据的加和结果, 此时也即实现了深度神经网络。

[0076] 以下对本发明具有多个闪存阵列结构单元的编码型闪存装置的编码方法作进一步的解释。

[0077] 根据本发明的实施例, 其中, 编码运算为全连接运算或卷积层运算。

[0078] 对于全连接层运算, 如图 3 所示, 输入向量 $X (1 \times M)$ 与权值矩阵向量 $K (M \times N)$ 相乘后得到输出向量 $Y (1 \times N)$, 数学表达如下公式 (1):

$$[0079] \quad Y_i = X_1 \cdot K_{1,i} + X_2 \cdot K_{2,i} + \dots + X_M \cdot K_{M,i}$$

[0080] 其中 $1 \leq i \leq N$ 。

[0081] 对于卷积层运算, 如图 5 所示, 多个局部区域的全连接层, 输入向量的矩阵 $X (M \times N)$ 的各部分分别与卷积矩阵 $K (k \times k)$ 进行矩阵向量乘法运算, 完成卷积操作并得到输出向量的矩阵 $Y (m \times n)$, 数学表达如下公式 (2):

$$[0082] \quad Y_{i,j} = X_{i,j} \cdot K_{k,k} + X_{i,j+1} \cdot K_{k,k-1} + X_{i+1,j} \cdot K_{k-1,k} + \dots + X_{i+k-1,j+k-1} \cdot K_{1,1}$$

[0083] 其中 $1 \leq i \leq M-k+1 (m)$, $1 \leq j \leq N-k+1 (n)$ 。

[0084] 根据本发明的又一实施例, 如图 3 和图 4 所示, 其中, 本发明实施例的编码方法基于图 3 所示的编码闪存装置实现, 编码运算为全连接运算时, 在步骤 S410 之前, 编码方法还包括:

[0085] 步骤 S510: 将全连接层的权值矩阵向量中的多个非零位元素中的每个非零位元素预存入编码型闪存装置中多个运算单元 C 中的对应运算单元 C 中。

[0086] 根据本发明的又一实施例,其中,在步骤S510之后,编码方法还包括:

[0087] 步骤S520:将对应的输入向量的输入元素对应输入多条字线中的每条字线中,以生成相应每条字线上的字线电压,使得编码型闪存装置的多个冗余单元处于开态。

[0088] 根据本发明的又一实施例,其中,在步骤S520之后,还包括:

[0089] 步骤S530:对编码型闪存装置的串选择线和地选择线施加高电平,以生成其中每个闪存阵列结构单元的多条源线中每条源线上的源线电压。

[0090] 具体地,对于编码运算为全连接运算时,如图3和图4所示,该编码型闪存装置中的一个闪存阵列结构单元100(即3D NAND FLASH结构单元)可以作为单个“串(serial)”负责处理单比特输入向量 $X(1 \times M)$ 与单比特权值矩阵 $K(M \times N)$ 的全连接运算。

[0091] 同一个串上共有N组存算模块,分别对应输出向量 $Y-1 \sim Y-N$ 。每组存算模块包含M条源线SL,分别对应输入字线 $WL-1 \sim WL-M$ 。运算开始之前,需对3D NAND FLASH结构单元的存算资源进行配置,按照进行向量矩阵乘法时输入向量与权值矩阵向量的各元素对应关系将权值矩阵的非零位元素写入3D NAND FLASH结构单元中对应的运算单元C,即对应于步骤S510。

[0092] 不同串的输入向量($X_1 \sim X_M$)的输入元素施加于对应的字线 $WL1 \sim WL-M$ 上,分别转化为对应 $WL1 \sim WL-M$ 上的输入字线电压,对应于步骤520,使得3D NAND FLASH结构单元上的冗余单元处于开态。

[0093] 当串选择线SSL线和地选择线GSL线加高电平时,编码运算开始:因为同一条源线SL只有一个FLASH单元(即运算单元C),该FLASH单元存储权重值(即前述的权值矩阵的非零位元素)并进行运算,其运算结果(0或1)可通过源线电压反映,即对应于步骤S530。

[0094] 该源线电压对应的源线电流值被比较器读出并转换为二进制形式的输出结果。N个加法器分别将对应的N组比较器的输出结果(每组比较器对应M条SL)的运算结果相加,即可得到全连接层的输出结果 $Y1 \sim Y-N$ 。

[0095] 根据本发明的又一实施例,不同串($SSL-1 \sim SSL-N$)协同处理多比特输入和多比特权值矩阵的全连接运算。其中,编码运算为全连接运算时,其中,输入向量为多比特数据,多比特数据可以是多比特输入向量和多比特权值矩阵向量的数据。步骤S410包括:

[0096] 基于多个闪存阵列结构单元同时进行对应输入向量和权值矩阵向量的编码运算,由于多个闪存阵列结构单元在本发明的编码型闪存装置中为并联连接,因此,其可以同时实现对多比特输入和多比特权值矩阵的全连接运算,即多个闪存阵列结构单元的每个闪存阵列结构单元实现单比特数据的全连接运算。

[0097] 或者基于一个闪存阵列结构单元采用分时复用的方式进行对应输入向量和权值矩阵向量的编码运算;对于闪存阵列结构单元的数量较为有限的情况,可以基于一个闪存阵列结构单元在不同的时间,分为多次进行多比特输入和多比特权值矩阵的全连接运算,即每次该闪存阵列结构单元均实现单比特数据的全连接运算。

[0098] 可见,本发明实施例的编码方法,基于3D NAND FLASH结构单元的存内计算架构可在一个时钟周期内完成并行的全连接运算。

[0099] 根据本发明的另一实施例,如图5和图6所示,其中,本发明实施例的编码方法基于图5所示的编码闪存装置实现,编码运算为卷积层运算时,在步骤S410之前,编码方法还包括:

[0100] 步骤S610:将卷积层的卷积矩阵中的多个卷积核元素中的每个卷积核元素预存入编码型闪存装置中多个运算单元中的对应运算单元中。

[0101] 根据本发明的另一实施例,其中,在步骤S610之后,编码方法还包括:

[0102] 步骤S620:将对应的输入向量的输入元素对应输入多条字线中的每条字线中,以生成相应每条字线上的字线电压,使得编码型闪存装置的多个冗余单元处于开态。

[0103] 根据本发明的另一实施例,其中,在步骤S620之后,编码方法还包括:

[0104] 步骤S630:对编码型闪存装置的串选择线和地选择线施加高电平,以生成其中每个闪存阵列结构单元的多条源线中每条源线上的源线电压。

[0105] 具体地,对于编码运算为卷积层运算时,如图5和图6所示,3D NAND FLASH结构单元的每个串负责处理单比特输入矩阵 $X(M \times N)$ 与卷积核 $K(k \times k)$ 的卷积运算。以附图5举例,输入 3×3 图像 $(X_{1,1}, X_{1,2}, X_{1,3}, X_{2,1}, X_{2,2}, X_{2,3}, X_{3,1}, X_{3,2}, X_{3,3})$ 与 2×2 卷积核 $(K_{1,1}, K_{1,0}, K_{0,1}, K_{0,0})$ 卷积后得到的 2×2 输出图像为 $(Y_{1,1}, Y_{1,2}, Y_{2,1}, Y_{2,2})$ 。实现上述卷积运算,输入图像(即对应的输入矩阵向量)需拆分成四个子矩阵 $(X_{1,1}, X_{1,2}, X_{2,1}, X_{2,2})$, $(X_{1,2}, X_{1,3}, X_{2,2}, X_{2,3})$, $(X_{2,1}, X_{2,2}, X_{3,1}, X_{3,2})$, $(X_{2,2}, X_{2,3}, X_{3,2}, X_{3,3})$ 分别与卷积核 $(K_{1,1}, K_{1,0}, K_{0,1}, K_{0,0})$ 向量矩阵乘法得到 $Y_{1,1}$, $Y_{1,2}, Y_{2,1}, Y_{2,2}$ 。

[0106] 在编码运算时,卷积层的卷积矩阵中每个卷积核元素被映射到输入向量 $(X_{1,1}, X_{1,2}, X_{2,1}, X_{2,2})$ 的对应运算单元C中,即位线WL1, WL2, WL4, WL5与各源线SL对应的交叉位置,即对应于步骤S610。考虑到神经网络中每个卷积层均有多个卷积核对输入图像进行多维度特征提取,因此可以将其他卷积核按照相同的原理分别写入WL1~WL9的对应运算单元C中。

[0107] 输入图像沿行的方向拆分成行向量 $(X_{1,1}, X_{1,2}, \dots, X_{1,N}, \dots, X_{M,1}, X_{M,2}, \dots, X_{M,N})$ 并施加到3D NAND FLASH结构单元的字线WL上作为输入,以分别转化为对应字线WL上的输入字线电压,对应于步骤S620。

[0108] 当串选择线SSL线和地选择线GSL线加高电平时,运算开始:因为同一条源线SL只有一个FLASH单元(即运算单元C),该FLASH单元存储卷积核元素值并进行运算,其运算结果(0或1)可通过源线电压反映,即对应于步骤S630。最后,如图5所示,4个加法器300分别将对应的4组比较器200的输出结果(每组4条源线SL)的输出结果相加,即可得到卷积层运算的输出结果 $(Y_{1,1}, Y_{1,2}, Y_{2,1}, Y_{2,2})$,此即实现了神经网络。

[0109] 根据本发明的另一实施例,其中,编码运算为卷积层运算时,其中,卷积层的卷积矩阵为多比特数据。本发明的编码方法还包括:

[0110] 步骤S440:对多个加法器的加和运算输出的加和结果进行移位求和,实现神经网络;具体地,当卷积核为多比特数据时,将编码运算得到的加法器的加和结果进行移位求和,即可以得到输入图像与多比特卷积核卷积后的输出图像。

[0111] 可见,本发明的编码方法基于上述编码闪存装置,可以在基于3D NAND FLASH结构的存内计算架构上,实现在一个时钟周期内完成并行的卷积运算。

[0112] 可见,在本发明的编码方法中,编码运算开始前,权值矩阵向量中的非零位元素或卷积向量中的卷积核元素依据一定的映射规则写入3D NAND FLASH结构单元对应的位置(运算单元),输入向量中不同元素的值(0或1)决定闪存阵列结构单元中字线WL上的输入电压(即字线电压 $V_g = V_{g_High}$ 或 V_{g_Low}),然后对应的源线SL的输出源线电流反映运算单元的计

算结果。最后,通过在3D NAND FLASH结构单元中各条源线SL的末端设置的多个比较器,将计算结果转换为二进制形式(0或1)并行读出。在此基础上,与比较器连接的加法器对来自多条源线的比较器的输出结果进行加和运算,得到最终的加和运算结果,实现高效且精确的全连接或卷积层运算,从而实现深度神经网络。

[0113] 至此,已经结合附图对本发明实施例进行了详细描述。

[0114] 需要说明的是,在附图或说明书正文中,未绘示或描述的实现方式,均为所属技术领域普通技术人员所知的形式,并未进行详细说明。此外,上述对各元件和方法的定义并不仅限于实施例中提到的各种具体结构、形状或方式,本领域普通技术人员可对其进行简单地更改或替换。

[0115] 还需要说明的是,实施例中提到的方向用语,例如“上”、“下”、“前”、“后”、“左”、“右”等,仅是参考附图的方向,并非用来限制本发明的保护范围。贯穿附图,相同的元素由相同或相近的附图标记来表示。在可能导致对本发明的理解造成混淆时,将省略常规结构或构造。

[0116] 并且图中各部件的形状和尺寸不反映真实大小和比例,而仅示意本发明实施例的内容。另外,在权利要求中,不应将位于括号之间的任何参考符号构造成对权利要求的限制。

[0117] 再者,单词“包含”不排除存在未列在权利要求中的元件或步骤。位于元件之前的单词“一”或“一个”不排除存在多个这样的元件。

[0118] 说明书与权利要求中所使用的序数例如“第一”、“第二”、“第三”等的用词,以修饰相应的元件,其本身并不意味着该元件有任何的序数,也不代表某一元件与另一元件的顺序或是制造方法上的顺序,这些序数的使用仅用来使具有某命名的一元件得以和另一具有相同命名的元件能做出清楚区分。

[0119] 本领域那些技术人员可以理解,可以对实施例中的设备中的模块进行自适应性地改变并且把他们设置在与该实施例不同的一个或多个设备中。可以把实施例中的模块或单元或组件组合成一个模块或单元或组件,以及此外可以把他们分成多个子模块或子单元或子组件。除了这样的特征和/或过程或者单元中的至少一些是相互排斥之外,可以采用任何组合对本说明书(包括伴随的权利要求、摘要和附图)中公开的所有特征以及如此公开的任何方法或者设备的所有过程或单元进行组合。除非另外明确陈述,本说明书(包括伴随的权利要求、摘要和附图)中公开的每个特征可以由提供相同、等同或相似目的的代替特征来代替。并且,在列举了若干装置的单元权利要求中,这些装置中的若干个可以是通过同一个硬件项来具体体现。

[0120] 类似地,应当理解,为了精简本发明并帮助理解各个公开方面的一个或多个,在上面对本发明的示例性实施例的描述中,本发明的各个特征有时被一起分组到单个实施例、图、或者对其的描述中。然而,并不应将该公开的方法解释成反映如下意图:即所要求保护的本发明要求比在每个权利要求中所明确记载的特征更多的特征。更确切地说,如下面的权利要求书所反映的那样,公开方面在于少于前面公开的单个实施例的所有特征。因此,遵循具体实施方式的权利要求书由此明确地并入该具体实施方式,其中每个权利要求本身都作为本发明的单独实施例。

[0121] 以上所述的具体实施例,对本发明的目的、技术方案和有益效果进行了进一步详

细说明,所应理解的是,以上所述仅为本发明的具体实施例而已,并不用于限制本发明,凡在本发明的精神和原则之内,所做的任何修改、等同替换、改进等,均应包含在本发明的保护范围之内。

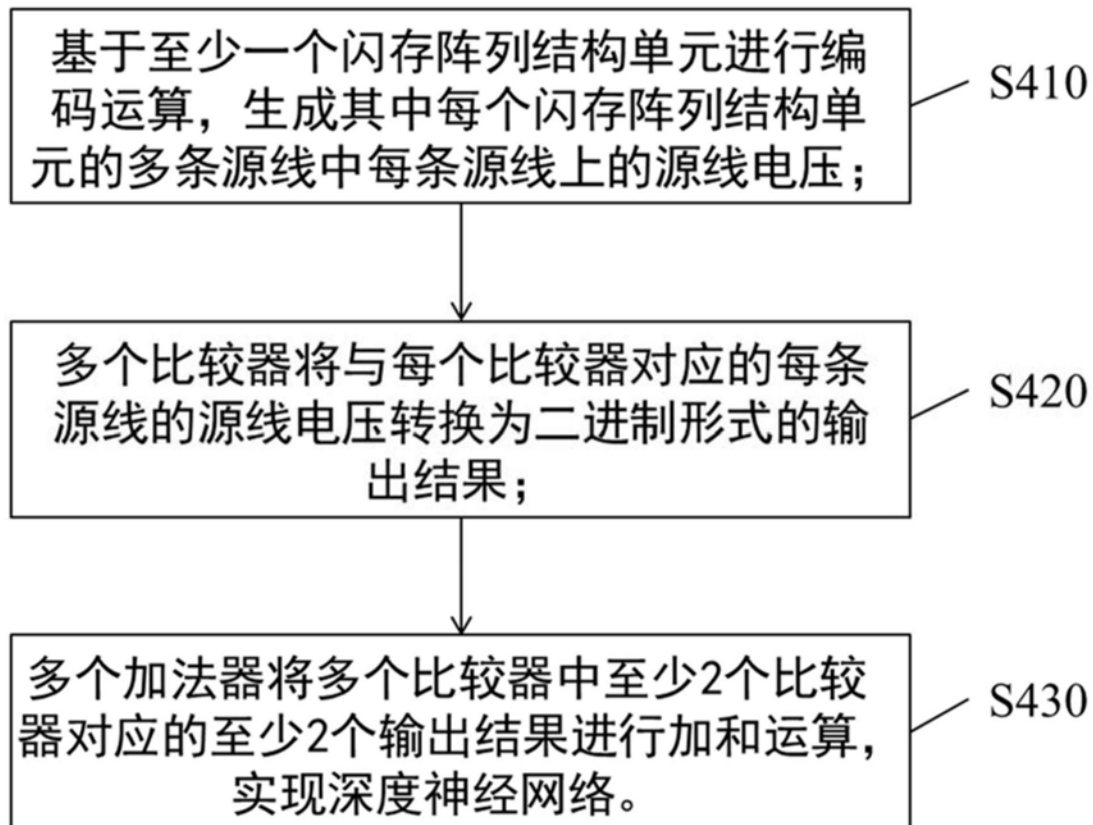


图2

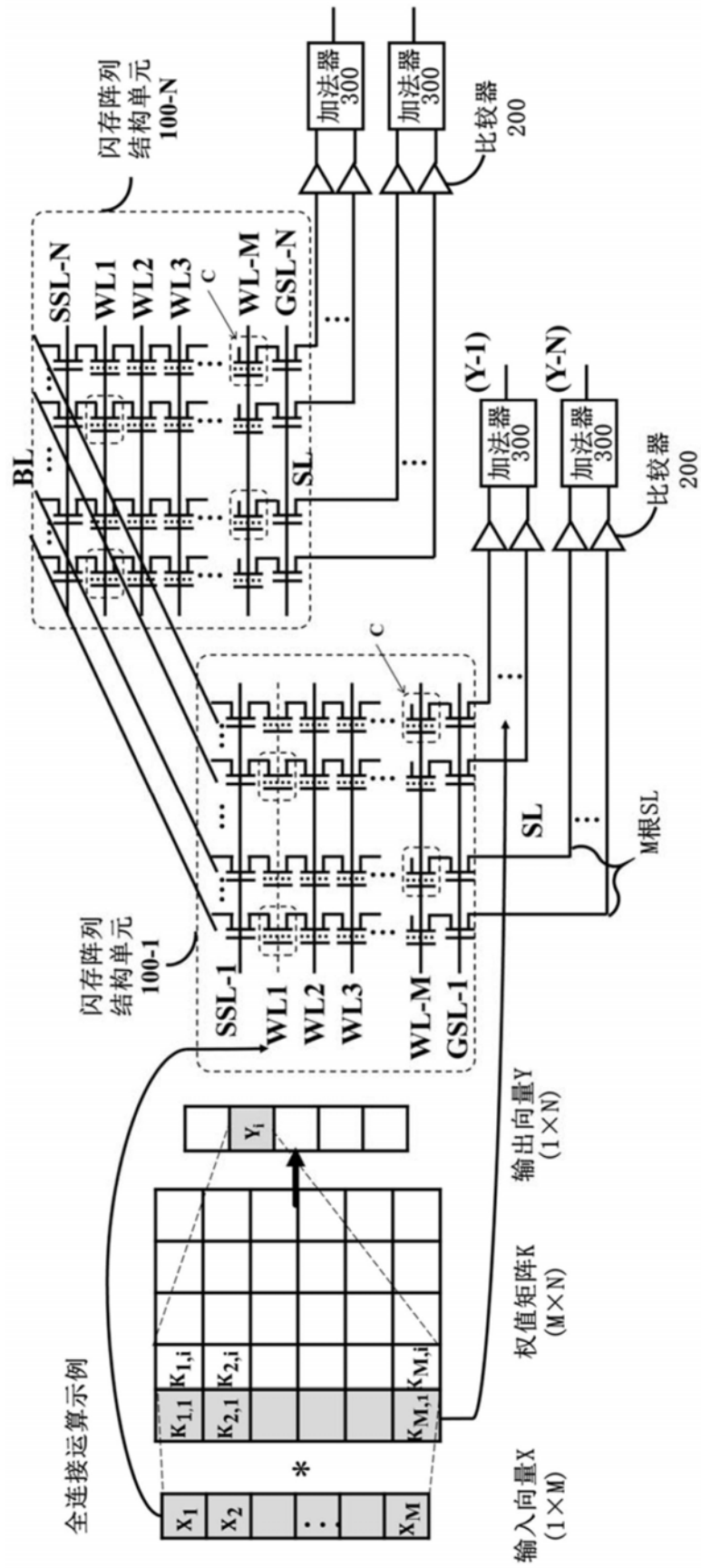


图3

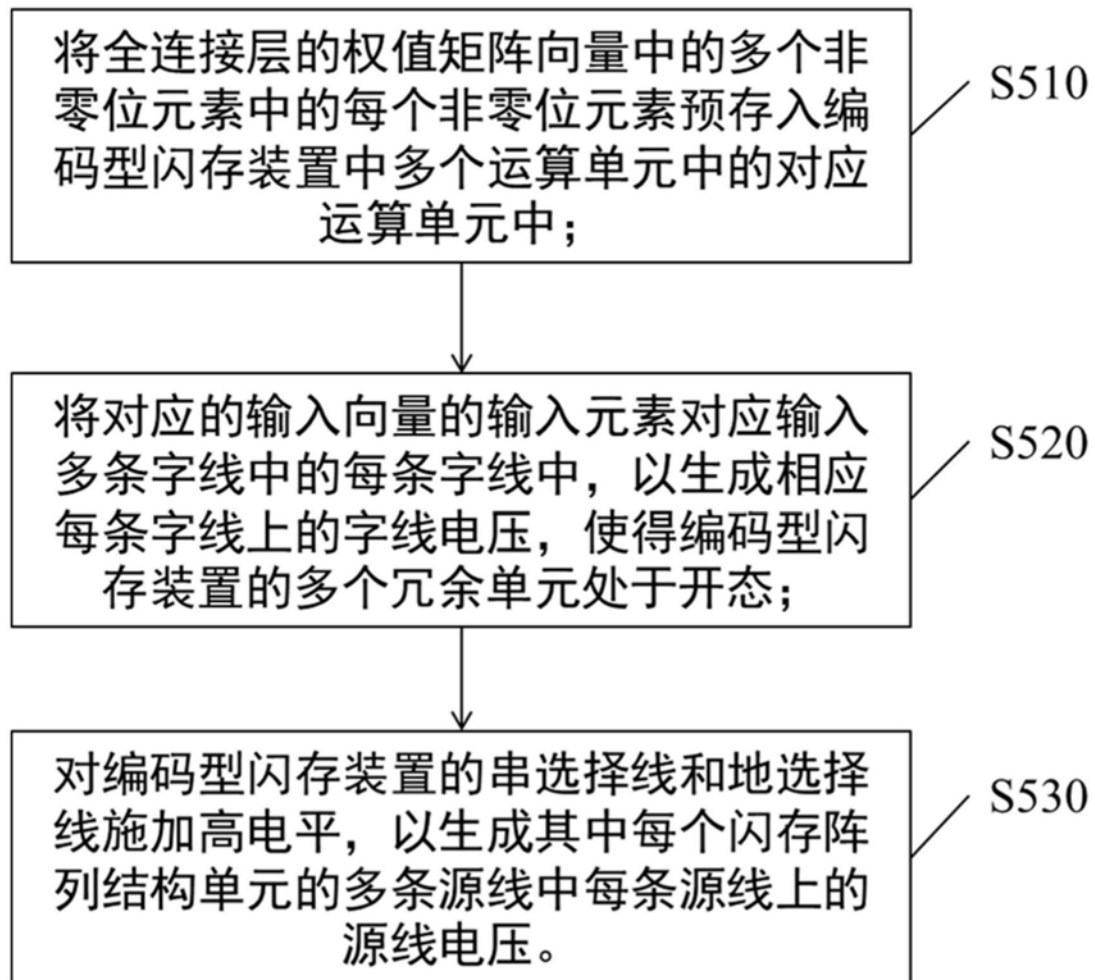


图4

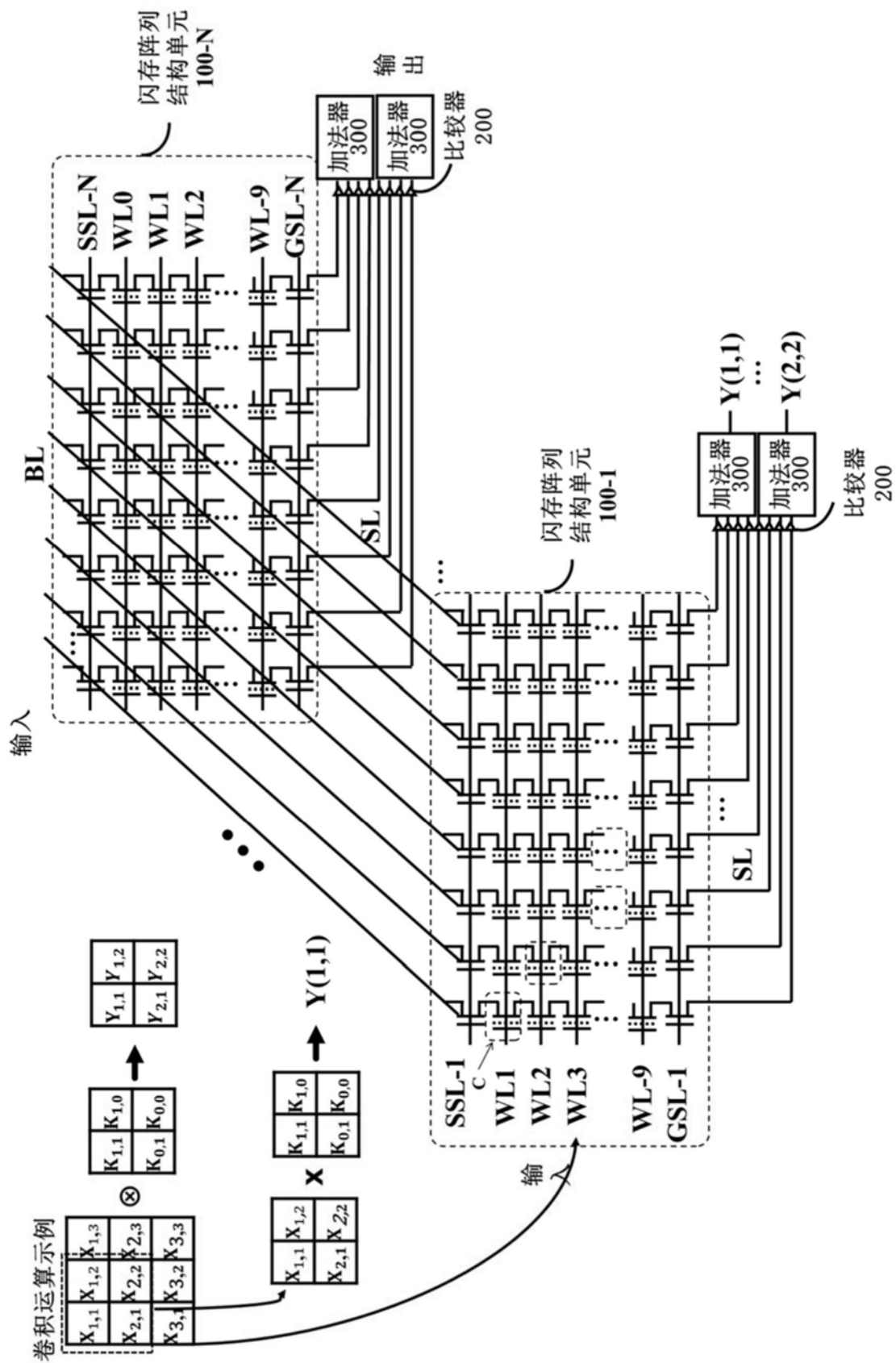


图5

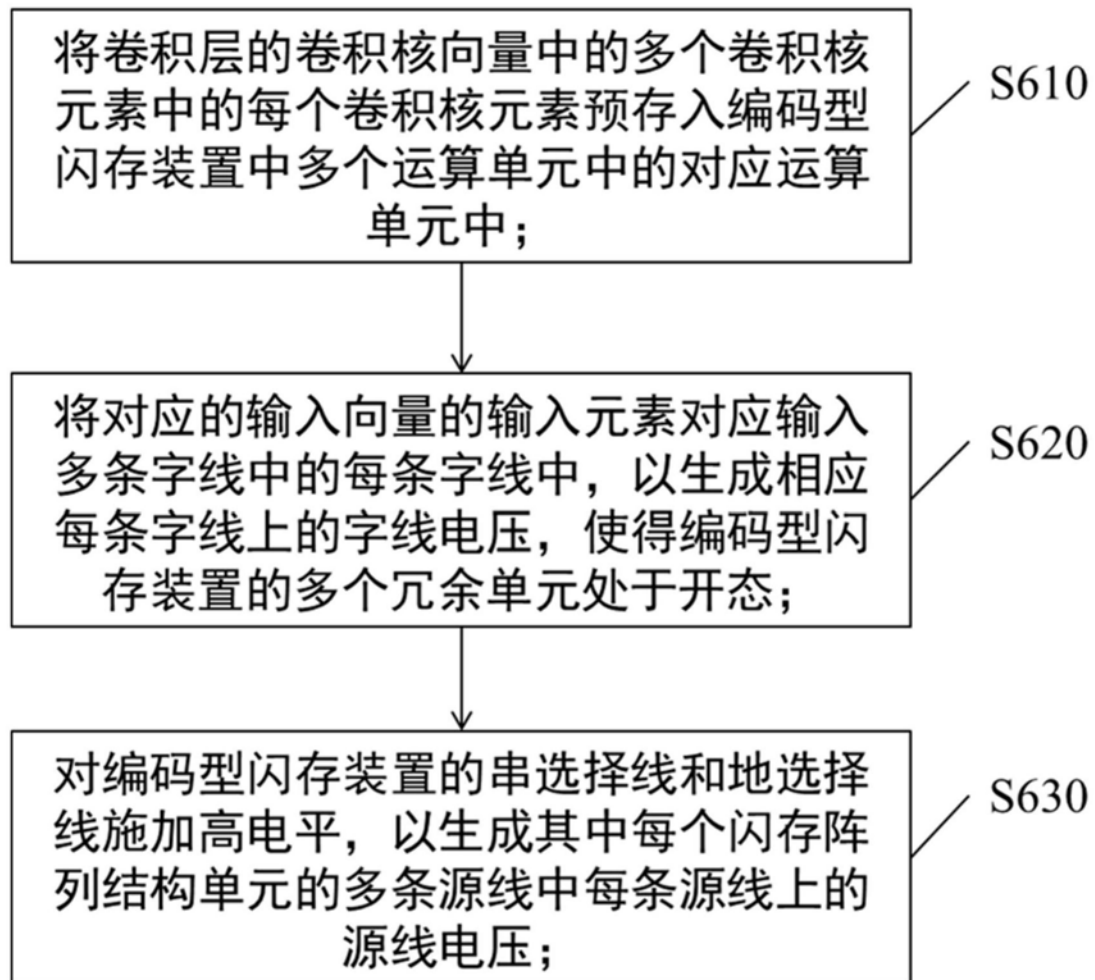


图6