(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 119207518 B
(45)授权公告日 2025.06.20

(21)申请号 202410001792.8
(22)申请日 2024.01.02
(65)同一申请的已公布的文献号
申请公布号 CN 119207518 A
(43)申请公布日 2024.12.27
(73)专利权人 百代(上海)数据技术有限公司
地址 200000 上海市长宁区定西路1328号3楼336室
(72)发明人 孙楠楠
(74)专利代理机构 无锡知之火专利代理事务所(特殊普通合伙) 32318
专利代理师 袁粉兰
(51)Int.Cl.
G11C 16/10(2006.01)
G11C 16/24(2006.01)
G06N 3/063(2023.01)

(56)对比文件
CN 110543937 A,2019.12.06

审查员 魏婷婷

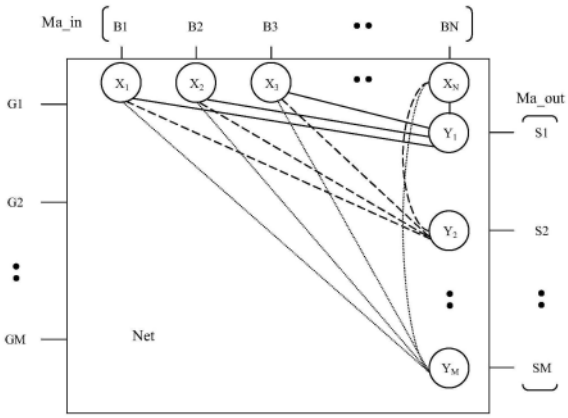
权利要求书2页 说明书15页 附图5页

(54)发明名称

执行存内计算的方法、装置及存储系统

(57)摘要

本发明涉及到执行存内计算的方法、装置及存储系统。在执行存内计算的闪存存储器当中利用存储晶体管阵列作为神经网络的权重矩阵之载体。设定神经网络的输入数据输送至存储晶体管阵列的一系列位线、以及设定从存储晶体管阵列的一系列源线处撷取到该神经网络的输出数据。其中,在存储晶体管阵列中任一晶体管的权重之大小利用其阈值电压进行调节,藉此来动态配置神经网络的权重矩阵。



CN 119207518 B

1.一种执行存内计算的方法,其特征在于,包括:

在闪存存储器中利用存储晶体管阵列作为神经网络的权重矩阵之载体,设定神经网络的输入数据输送至存储晶体管阵列的一系列位线、以及设定从存储晶体管阵列的一系列源线处撷取到该神经网络的输出数据;其中,存储晶体管阵列中任一晶体管的权重之大小利用其阈值电压进行调节,藉此来动态配置神经网络的权重矩阵;

针对存储晶体管阵列的任一字线上的任一晶体管在动态配置权重的阶段,该任一晶体管在执行任意一次权重调节之前,先将该任一字线上相邻前一次权重调节的其他晶体管所使用的编程电压的反向电压瞬态挂载一次到该任一字线上、再将为该任一晶体管所设定的编程电压施加到该任一字线上以执行该任意一次权重调节,从而抑制存储晶体管阵列的该任一字线处存在的寄生电容的寄生电荷所导致的权重更新迭代失败。

2.根据权利要求1所述的方法,其特征在于:

用于神经网络的所述存储晶体管阵列位于闪存存储器的同一个物理块上;以及

神经网络的输入数据、输出数据各自所在的存储空间与所述存储晶体管阵列所在的存储空间分别属于闪存存储器的不同物理块。

3.根据权利要求1所述的方法,其特征在于:

神经网络的原始的输入数据转换成电压值形式的输入量,以基于排序成向量的输入量输送至存储晶体管阵列的一系列位线。

4.根据权利要求1所述的方法,其特征在于:

神经网络的输出数据以存储晶体管阵列的输出电流表征,存储晶体管阵列的一系列源线的输出电流被视为存储晶体管阵列的排序成向量的输出量。

5.根据权利要求1所述的方法,其特征在于:

调节任一晶体管的权重的方式包括改变其浮栅注入电荷,藉由晶体管的浮栅注入电荷的改变而调节任一晶体管在存储晶体管阵列中所处位置处的权重大小。

6.根据权利要求5所述的方法,其特征在于:

在神经网络的反向传播阶段,通过重新设置存储晶体管阵列中各个位置处的权重大小而动态配置神经网络的权重矩阵。

7.根据权利要求3所述的方法,其特征在于:

在存储晶体管阵列的每个位线处皆配置有数模转换器,其用于将神经网络的输入数据转换成与位线的输入信号模式相匹配的电压值。

8.根据权利要求4所述的方法,其特征在于:

在存储晶体管阵列的每个源线处皆配置有模数转换器,其用于将源线处的输出电流格式的输出信号转换成与神经网络的数字输出量相适配的输出数据。

9.一种执行存内计算的装置,其特征在于,包括:

闪存存储器,由其存储晶体管阵列作为神经网络的权重矩阵之载体,设定神经网络的输入数据输送至存储晶体管阵列的一系列位线、以及设定从存储晶体管阵列的一系列源线处撷取到该神经网络的输出数据,其中,存储晶体管阵列中任一晶体管的权重之大小利用其阈值电压进行调节,藉此来动态配置神经网络的权重矩阵,针对存储晶体管阵列的任一字线上的任一晶体管在动态配置权重的阶段,该任一晶体管在执行任意一次权重调节之前,先将该任一字线上相邻前一次权重调节的其他晶体管所使用的编程电压的反向电压瞬

态挂载一次到该任一字线上、再将为该任一晶体管所设定的编程电压施加到该任一字线上以执行该任意一次权重调节,从而抑制存储晶体管阵列的该任一字线处存在的寄生电容的寄生电荷所导致的权重更新迭代失败;

输入单元,其将神经网络的原始的输入数据转换成电压值形式的输入量,以基于排序成向量的输入量输送至存储晶体管阵列的一系列位线;

输出单元,其用于将源线处的输出电流格式的输出信号转换成与神经网络的数字输出量相适配的输出数据,存储晶体管阵列的一系列源线的输出电流被视为存储晶体管阵列的排序成向量的输出量。

10.一种执行存内计算的存储系统,其特征在于,包括:

基于闪存存储器的存储芯片,在闪存存储器中利用存储晶体管阵列作为神经网络的权重矩阵之载体,其中任一晶体管的权重之大小利用其阈值电压进行调节,藉此来动态配置神经网络的权重矩阵;设定神经网络的输入数据输送至存储晶体管阵列的一系列位线以及还设定从存储晶体管阵列的一系列源线处撷取到该神经网络的输出数据,针对存储晶体管阵列的任一字线上的任一晶体管在动态配置权重的阶段,该任一晶体管在执行任意一次权重调节之前,先将该任一字线上相邻前一次权重调节的其他晶体管所使用的编程电压的反向电压瞬态挂载一次到该任一字线上、再将为该任一晶体管所设定的编程电压施加到该任一字线上以执行该任意一次权重调节,从而抑制存储晶体管阵列的该任一字线处存在的寄生电容的寄生电荷所导致的权重更新迭代失败;

内置在存储芯片中的数模转换器,其中在存储晶体管阵列的每个位线处皆配置有数模转换器,以将神经网络的输入数据转换成与位线的输入信号模式相匹配的电压值;

内置在存储芯片中的模数转换器,其中在存储晶体管阵列的每个源线处皆配置有模数转换器,以将源线处的输出电流格式的输出信号转换成与神经网络的数字输出量相适配的输出数据。

执行存内计算的方法、装置及存储系统

技术领域

[0001] 本发明主要涉及到数据存储的技术领域,确切的说,涉及到了在数据存储领域的可执行存内计算的方法和可执行存内计算装置以及相应的存储系统。

背景技术

[0002] 以往的电子设备中计算和存储是分开的,计算部分例如处理器和存储部分例如硬盘或内存等存储器各自分别专注于本职任务,计算和存储基本互不干涉对方。亦有厂商为了提高计算速度和计算能力而将处理器和存储器在物理位置上集成得比较近,例如在电路板的版图设计上它们之间相互靠拢、短线路、在封装阶段直接相互层叠布置。

[0003] 业界的算力提升一直主要是按照摩尔定律的节奏逐步推进,但主流的计算范式始终遵循着冯诺依曼架构。存算一体作为一种新型的算力,业界主要关注狭义的存算一体也即利用存储介质进行计算但计算媒介方法尚未完全统一。存算一体作为新型算力,有望解决传统冯诺依曼架构下的编译墙、存储墙、功耗墙问题。存算一体在业界的共识,其基本已被确定为解决算力问题的关键技术之一。

[0004] 存算一体将存储与计算有机融合,基于其巨大能效比提升潜力并有望成为算力时代的先进计算架构方案。但是存在着疑虑:如何将存储器映射到神经网络以及如何给神经网络配置可动态调节的权重矩阵,以适配于神经网络的实际应用例如反向传播等,这是采用存内计算必须处理的关键问题。尤其是:如何应对存储器内部神经网络权重更新时所面临的权重偏移等亟待解决的疑虑,对存储器而言这属于一种隐匿的难题。

发明内容

[0005] 本申请涉及一种执行存内计算的方法,其特征在于,包括:

[0006] 在闪存存储器中利用存储晶体管阵列作为神经网络的权重矩阵之载体,设定神经网络的输入数据输送至存储晶体管阵列的一系列位线、以及设定从存储晶体管阵列的一系列源线处撷取到该神经网络的输出数据;其中,存储晶体管阵列中任一晶体管的权重之大小利用其阈值电压进行调节,藉此来动态配置神经网络的权重矩阵。

[0007] 上述的方法,其中用于神经网络的所述存储晶体管阵列位于闪存存储器的同一个物理块上;以及神经网络的输入数据、输出数据各自所在的存储空间与所述存储晶体管阵列所在的存储空间分别属于闪存存储器的不同物理块。

[0008] 上述的方法,神经网络的原始的输入数据转换成电压值形式的输入量,以基于排序成向量的输入量输送至存储晶体管阵列的一系列位线。

[0009] 上述的方法,神经网络的输出数据以存储晶体管阵列的输出电流表征,存储晶体管阵列的一系列源线的输出电流被视为存储晶体管阵列的排序成向量的输出量。

[0010] 上述的方法,调节任一晶体管的权重的方式包括改变其浮栅注入电荷,藉由晶体管的浮栅注入电荷的改变而调节任一晶体管在存储晶体管阵列中所处位置处(也即权重矩阵中该任一晶体管对应的一个元素位置处)的权重大小。

[0011] 上述的方法,在神经网络的反向传播(Back propagation)阶段,通过重新设置存储晶体管阵列中各个位置处的权重大小而动态配置神经网络的权重矩阵。

[0012] 上述的方法,在存储晶体管阵列的每个位线处皆配置有数模转换器,其用于将神经网络的输入数据转换成与位线的输入信号模式相匹配的电压值。

[0013] 上述的方法,在存储晶体管阵列的每个源线处皆配置有模数转换器,其用于将源线处的输出电流格式的输出信号转换成与神经网络的数字输出量相适配的输出数据。

[0014] 本申请涉及一种执行存内计算的装置,其特征在于,包括:

[0015] 闪存存储器,由其存储晶体管阵列作为神经网络的权重矩阵之载体,设定神经网络的输入数据输送至存储晶体管阵列的一系列位线、以及设定从存储晶体管阵列的一系列源线处撷取到该神经网络的输出数据;

[0016] 输入单元,其将神经网络的原始的输入数据转换成电压值形式的输入量,以基于排序成向量的输入量输送至存储晶体管阵列的一系列位线;

[0017] 输出单元,其用于将源线处的输出电流格式的输出信号转换成与神经网络的数字输出量相适配的输出数据,存储晶体管阵列的一系列源线的输出电流被视为存储晶体管阵列的排序成向量的输出量。

[0018] 上述的执行存内计算的装置,其中用于神经网络的所述存储晶体管阵列位于闪存存储器的同一个物理块上;以及神经网络的输入数据、输出数据各自所在的存储空间与所述存储晶体管阵列所在的存储空间分别属于闪存存储器的不同物理块。

[0019] 上述的执行存内计算的装置,其中神经网络的原始的输入数据转换成电压值形式的输入量,以基于排序成向量的输入量输送至存储晶体管阵列的一系列位线。

[0020] 上述的执行存内计算的装置,其中神经网络的输出数据以存储晶体管阵列的输出电流表征,存储晶体管阵列(transistor arrays)的一系列源线的输出电流被视为存储晶体管阵列的排序成向量的输出量。

[0021] 上述的执行存内计算的装置,其中调节任一晶体管的权重的方式包括改变其浮栅注入电荷,藉由晶体管的浮栅注入电荷的改变而调节该任一晶体管在存储晶体管阵列中所处位置处(也即权重矩阵中该任一晶体管对应的一个元素位置处)的权重大小。

[0022] 上述的执行存内计算的装置,神经网络的反向传播(Back propagation)阶段可通过重新设置存储晶体管阵列中各个位置处的权重大小而动态配置神经网络的权重矩阵。

[0023] 上述的执行存内计算的装置,输入单元至少包括了数模转换器,在存储晶体管阵列的每个位线处皆配置有数模转换器,其用于将神经网络的输入数据转换成与位线的输入信号模式相匹配的电压值。

[0024] 上述的执行存内计算的装置,输出单元至少包括了模数转换器,在存储晶体管阵列的每个源线处皆配置有模数转换器,其用于将源线处的输出电流格式的输出信号转换成与神经网络的数字输出量相适配的输出数据。

[0025] 本申请涉及一种执行存内计算的存储系统,其特征在于,包括:

[0026] 基于闪存存储器的存储芯片,在闪存存储器中利用存储晶体管阵列作为神经网络的权重矩阵之载体,其中任一晶体管的权重之大小利用其阈值电压进行调节,藉此来动态配置神经网络的权重矩阵;设定神经网络的输入数据输送至存储晶体管阵列的一系列位线以及还设定从存储晶体管阵列的一系列源线处撷取到该神经网络的输出数据;

[0027] 内置在存储芯片中的数模转换器,其中在存储晶体管阵列的每个位线处皆配置有数模转换器,以将神经网络的输入数据转换成与位线的输入信号模式相匹配的电压值;

[0028] 内置在存储芯片中的模数转换器,其中在存储晶体管阵列的每个源线处皆配置有模数转换器,以将源线处的输出电流格式的输出信号转换成与神经网络的数字输出量相适配的输出数据。

[0029] 上述的存储系统,其中用于神经网络的所述存储晶体管阵列位于闪存存储器的同一个物理块上;以及神经网络的输入数据、输出数据各自所在的存储空间与所述存储晶体管阵列所在的存储空间分别属于闪存存储器的不同物理块。

[0030] 上述的存储系统,神经网络的原始的输入数据转换成电压值形式的输入量,以基于排序成向量的输入量输送至存储晶体管阵列的一系列位线。

[0031] 上述的存储系统,神经网络的输出数据以存储晶体管阵列的输出电流表征,存储晶体管阵列的一系列源线的输出电流被视为存储晶体管阵列的排序成向量的输出量。

[0032] 上述的存储系统,调节任一晶体管的权重的方式包括改变其浮栅注入电荷,藉由晶体管的浮栅注入电荷的改变而调节任一晶体管在存储晶体管阵列中所处位置处(也即权重矩阵中该任一晶体管对应的一个元素位置处)的权重大小。

[0033] 上述的存储系统,在神经网络的反向传播(Back propagation)阶段,通过重新设置存储晶体管阵列中各个位置处的权重大小而动态配置神经网络的权重矩阵。

[0034] 上述的存储系统,存储晶体管阵列的每个位线处皆配置有数模转换器,其用于将神经网络的输入数据转换成与位线的输入信号模式相匹配的电压值。

[0035] 上述的存储系统,存储晶体管阵列的每个源线处皆配置有模数转换器,其用于将源线处的输出电流格式的输出信号转换成与神经网络的数字输出量相适配的输出数据。

[0036] 本申请涉及一种执行存内计算的芯片,其特征在于,包括:

[0037] 存储晶体管阵列,由存储晶体管阵列作为神经网络的权重矩阵之载体,任一晶体管的权重之大小利用其阈值电压进行调节,藉此来动态配置神经网络的权重矩阵;设定神经网络的输入数据输送至存储晶体管阵列的一系列位线,以及设定从存储晶体管阵列的一系列源线处撷取到该神经网络的输出数据;

[0038] 集成在所述芯片中的数模转换器,其中在存储晶体管阵列的每个位线处皆配置有数模转换器,以将神经网络的输入数据转换成与位线的输入信号模式相匹配的电压值;

[0039] 集成在所述芯片中的模数转换器,其中在存储晶体管阵列的每个源线处皆配置有模数转换器,以将源线处的输出电流格式的输出信号转换成与神经网络的数字输出量相适配的输出数据。

[0040] 基于前文所述,本申请最大优势之一是将神经网络的权重矩阵直接映射到存储晶体管阵列并且可以通过对晶体管阈值电压进行调节,藉此来动态配置神经网络的权重矩阵以便适应于神经网络的权重矩阵(weight matrix)更新例如匹配反向传播。并藉此还可以应对存储器内部神经网络权重更新时所面临的权重偏移等较为棘手的难题,防止存储器所映射出的权重矩阵中的一个或多个元素发生非期望的隐藏偏移。

[0041] 基于前文所述,本申请最大优势之二是采用基于非冯诺依曼架构,可以解决传统冯诺依曼架构之下的编译墙、存储墙、功耗墙问题。在编译方面的一个便捷之处是通过调整晶体管的阈值电压而实现核心权重的编程,这种解决方案仅仅只需要在晶体管处配置晶体

管三端的电压或电流,即可实现核心权重类编译,因为是在传统存储器上进行兼容于电压电流等信息量的编译,所以在设计一部分编译行为方面不存在任何障碍。再者不像传统方案那样须要求存储器与处理器之间执行数据交换以及随之而来的存储墙,本申请数据存储和数据计算尤其是矩阵运算都发生在存储器内部,解决了存储墙的问题。由于神经网络的具体实现是基于无需在存储器与处理器之间实施大量数据搬运的前提下,而鉴于功耗墙的大部分能耗在数据搬运过程中产生例如数据搬运功耗是计算功耗的千倍以上,所以也还解决了神经网络及其人工智能应用场景下的高能耗难题。另外AI运算需1PB/s而数据搬运速度方面例如DRAM 40GB-1TB/s远远不足以满足AI运算的需求,如果使用本申请那样的存储器内部计算则可以较好解决数据搬运速度无法跟随AI运算速度的疑虑。

[0042] 基于前文所述,本申请最大优势之三是推进算力大幅度提升。通过使用存算一体技术则可将具AI计算的中大量乘加计算的权重部分存在存储单元中,在存储单元的已有存储电路部分上做修改,从而在读取的同时进行数据输入和计算处理,在存储阵列完成诸如卷积运算等神经网络。由于大量乘加的卷积运算是深度学习的核心组成部分,因此存内计算和存内逻辑非常适合人工智能的深度神经网络应用和基于AI的大数据技术。

附图说明

[0043] 为使上文目的和特征及优点能够更加通俗易懂,后文结合附图对具体实施方式做详细的阐释,阅读以下说明并参照以下附图之后,本申请的特征和优势将显而易见。

[0044] 图1是闪存存储器内部选取了一部分晶体管作为神经网络的权重载体。

[0045] 图2是从存储晶体管阵列的一系列位线处来输入神经网络的输入数据。

[0046] 图3是以图片或视频等为例将像素值或灰度值输入至存储晶体管阵列。

[0047] 图4是将神经网络的权重矩阵映射到闪存存储器的存储晶体管阵列处。

[0048] 图5是由浮栅注入电荷的改变而调节晶体管在晶体管阵列中的权重值。

[0049] 图6是应对存储器内部神经网络权重更新时所面临的权重偏移等疑虑。

具体实施方式

[0050] 下面将结合各具体的实施例,对本申请披露的技术方案进行清楚完整的阐述,但所描述的实施例仅是本申请用作叙述说明所用的实施例而非全部的实施例,在基于该等实施例的基础之上,本领域的技术人员应该认识到,在没有做出创造性劳动的前提下所获得的任何方案都属于本申请的保护范围。

[0051] 参见图1,这是一个已经存在的通用型NOR闪存存储器的电路,其工作原理是在存储晶体管的源极S和漏极D之间,也即电流单向传导的半导体上可以形成贮存电荷如电子的浮动栅(Floating Gate),浮动栅被包裹着一层硅氧化膜绝缘体。浮动栅上面是在源极和漏极之间控制传导电流的选择/控制栅极(Control Gate)。浮栅的电荷如电子不会因掉电而消失,所以闪存是非易失存储器。写操作如在控制极加正电压,使电荷如电子通过绝缘层进入浮栅极。擦除操作正好相反,如在衬底加正电压,将电子从浮栅极中析出来从而实现擦除。闪存的数据操作属于现有技术,本文不再对其单独赘述,闪存存储器在现有技术中的读或写或擦除等操作可适用于本申请的闪存存储器。

[0052] 参见图1,以阵列形式描述闪存存储器:每行晶体管的栅极连接到共同的字线以及

每列晶体管的漏极连到公共的位线、每列晶体管的源极连到公共的源线。

[0053] 参见图1,关于行字线范例,第一行晶体管各自的栅极连到字线G1以及第二行晶体管各自的栅极连到字线G2、第M行晶体管各自的栅极连到字线GM。更多的省略行未重复展示在图中。合计M行晶体管视为矩阵M×N的行元素。

[0054] 参见图1,关于行源线范例,第一行晶体管各自的源极连到源线S1以及第二行晶体管各自的源极连到源线S2、第M行晶体管各自的源极连到源线SM。更多的省略行未重复展示在图中。

[0055] 参见图1,关于列位线范例,第一列晶体管各自的漏极连到位线B1以及第二列晶体管各自的漏极连到位线B2、第N列晶体管各自的漏极连到位线BN。更多的省略列未重复展示在图中。合计N列晶体管视为矩阵M×N的列元素。

[0056] 参见图1,图示的闪存存储器原本的功能是保存存储码元1或0,换言之闪存存储器在传统计算机或服务器或其他类似的电子设备中,其并无特殊的用途。但是本申请不再将闪存存储器视为仅仅保存码元1或0的存储空间,而是主张将闪存存储器的电路架构改为可参与人工智能的类计算单元,闪存存储器的底层物理结构近乎不变。

[0057] 参见图2,利用基于闪存存储器的类计算单元来替代传统计算机,是因为计算机在执行神经网络的运行阶段,存在几大弊端:从处理器外的存储器读写数据,数据搬运时间是运算时间的成百上千倍,整个过程的无功能耗大约在60%-90%之间,能效非常低且顺带的存储墙这种问题成为了数据计算应用的一大障碍。尤其是,神经网络中深度学习加速的最大挑战就是数据在计算单元和存储单元之间频繁的移动。虽然例如多核CPU或者是例如众核GPU并行加速技术被用于提升算力,但后摩尔时代,存储带宽制约了计算系统的有效带宽并导致整体性算力的增长步履维艰。在传统计算机上运行神经网络的权重计算或者运行神经网络的推理计算,耗时耗电就是这种弊端的最突出表现之一。

[0058] 参见图2,神经网络引入大量参数使其发展遇到了瓶颈。全连接层每一个输出神经元都与所有输入神经元相连接。按照图中假设:若输入神经元的数量为N,输出神经元的数量为M,则至少大约有N×M个权重或者有M个偏置需要训练与计算。而卷积层对于每个输出神经元,通常也还有卷积核覆盖范围内的输入神经元与其相连接,且所有输出使用的卷积核一致,局部连接与权值共享虽然减少卷积层的参数,但是大尺寸图像输入与深度神经网络条件下仍然有大量计算需要被执行。大量数据的读取与计算面临着网络的内存瓶颈和运算瓶颈,神经网络的功耗和时延急剧上升,限制着神经网络的应用。

[0059] 参见图2,为了解决前述瓶颈,图示的非冯诺伊曼架构,在特定领域可以提供更大算力和更高能效,可超越现有ASIC算力芯片。存内计算和存内逻辑,即存算一体技术直接利用存储器进行数据处理或计算,从而将数据存储与数据计算融合在相同芯片或相同存储器的同一片区之中,可以彻底消除冯诺依曼计算架构的瓶颈,尤其是较适用于深度学习神经网络这种大数据量大规模并行的应用场景。

[0060] 参见图2,在设计中位线B1-BN作为输入、源线S1-SM作为输出,图1的闪存存储器视为可实现神经网络NET的物理载体。

[0061] 参见图2,为了阐释具体的实现过程,先设计一个如下的矩阵Ma_wei。

[0062] $$Ma_wei=\begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1N} \\ w_{21} & w_{22} & \cdots & w_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ w_{M1} & w_{M2} & \cdots & w_{MN} \end{bmatrix}$$

[0063] 参见图2,位线B1-BN输入以向量Ma_in＝[B1、B2…BN]表示。为了数学表达上的方便需将位线B1的输入量映射到X_1、将位线B2的输入量映射到X_2、将位线B3的输入量映射到X_3、……,将位线BN的输入量映射到X_N。

[0064] 参见图2,源线S1-SM输出以向量Ma_out＝[S1、S2…SM]表示。因数学表达上的方便需将源线S1的输出量映射到Y_1、将源线S2的输出量映射到Y_2、将源线S3的输出量映射到Y_3、……,将源线SM的输出量映射到Y_M。

[0065] 参见图2,闪存存储器之目的在于执行神经网络NET的运算功能,具体的运算过程用以下的式子来表达。实质上这个过程实际上也是将闪存存储器模拟成一个计算单元或类计算单元的详细范例。以Y_1和Y_2及Y_M等作为代表来阐明这一精神。

[0066] $Y_1＝X_1*W_{11}+X_2*W_{12}+X_3*W_{13}+\cdots X_N*W_{1N}$

[0067] $Y_2＝X_1*W_{21}+X_2*W_{22}+X_3*W_{23}+\cdots X_N*W_{2N}$

[0068] $Y_M＝X_1*W_{M1}+X_2*W_{M2}+X_3*W_{M3}+\cdots X_N*W_{MN}$

[0069] 参见图2,更直观的表达是利用向量矩阵乘法来展示前述复杂运算。向量矩阵表达闪存存储器执行神经网络NET的局部数据处理如下所述。

[0070] $$\begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1N} \\ w_{21} & w_{22} & \cdots & w_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ w_{M1} & w_{M2} & \cdots & w_{MN} \end{bmatrix} * \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_M \end{bmatrix}$$

[0071] 参见图2,如果定义X_1至X_N为神经网络NET的输入层,那么Y_1至Y_N对应就是神经网络NET的输出层。至此,可认为图1的闪存存储器完成了一次从输入层到输出层的整个向量运算,而且这种运算和传统冯诺伊曼计算机的运算的大为不同,因为其运算只要从源线S1-SM测得各自的信号量即可,单个乘法和加法无须一一实施,最终的运算结果能够一次性从源线S1-SM的信号量得以充分体现。所以称之为类计算。传统的冯诺伊曼计算机的运算则必须将所有的乘法和加法一一实施,这是两者的最大不同。

[0072] 参见图2,基于闪存存储器的神经网络NET适用于图像处理领域。假设带有像素值或者灰度值的图片或视频(人工智能、自动驾驶、人脸识别、目标检测)等的数据被输入至基于闪存存储器的神经网络NET,下文以图像Gr_in为例。

[0073] 参见图3,图像Gr_in中的像素点分为多行和多列,定义第0个像素的地址所对应的像素为g1、定义第1个像素的地址所对应的像素为g2、定义第2个像素的地址所对应的像素为g3,依此类推,末尾一个像素地址对应的像素为gn。按照像素矩阵的排列规律和地址规定,任意一个像素点的灰度值可以通过拍摄得到的图像而得到。允许n个像素点的灰度值或像素值(属于图像Gr_in)耦合至闪存存储器的X_1至X_N等,当然灰度值或像素值也还可以进行归一化处理或实施类似的图像数据处理。

[0074] 参见图3,值得关注的一个问题是,位线B1-BN的输入量格式对于闪存存储器而言

应当符合闪存存储器的内在特征,也即,位线B1-BN的输入值与像素值g1-gn并不是可直接对等的量。所以在可选的实施例中,像素值g1-gn经过DAC器件例如和闪存存储器封装到一起的数模转换器之数字量到模拟量的转换处理,像素值g1-gn各自对应的模拟量再输入至位线B1-BN,DAC(Digital to Analog Converter)器件也还可以和闪存存储器在晶圆制造阶段而直接集成到相同的半导体晶片上。至此,图像Gr_in蕴含的数据可以顺畅的与闪存存储器耦合。另一个重要的因素还在于:图像Gr_in其最终的保存位置或空间通常必然是各类存储器,那么图像Gr_in的保存位置以及执行矩阵运算的位置都是相同的闪存存储器这种内禀特性的突出优势便可显现出来,处理器CPU或GPU将不必再从存储器读取数据到自身而执行矩阵运算,存储动作和计算动作全部在存储器实施,前文中提及从处理器到存储器之间的读写数据、数据搬运时间成本等疑虑将不复存在。

[0075] 参见图3,有一个重要问题还没有解决:前文为了阐释神经网络的具体实现过程而特意设计了参与网络计算的矩阵Ma_wei。然而如何将M*N向量矩阵Ma_wei对偶到闪存存储器内晶体管的如图1所示的M×N的元素上仍然是亟待解决的疑虑。

[0076] 参见图4,关于像素值g1-gn与位线B1-BN的关系,像素值g1经DAC数模转换处理而折算成电压值V1、像素值g2经数模转换而折算成电压值V2、像素值g3经数模转换而折算成电压值V3,依此类推,像素值gn经数模转换而折算成电压值VN。按照图示的接口对应关系,电压值V1对接到位线B1、电压值V2对接到位线B2、电压值V3对接输入到位线B3,依此类推,电压值VN对接输入到位线BN。按照规律,像素值较大者其转换的位线电压值也较大,像素值较小者其转换的位线电压值反之较低。再者晶体管的布局数量也即矩阵M×N的大小应当与图像Gr_in的像素尺寸具有对应关系。

[0077] 参见图4,关于矩阵Ma_wei,第一行晶体管各自的栅极连到字线G1以及第二行晶体管各自的栅极连到字线G2、第M行晶体管各自的栅极连到字线GM。第一行晶体管在权重矩阵中第一行它们各自产生的权重记作W_{11}、W_{12}、W_{13}、……W_{1N}。第二行晶体管在权重矩阵中第二行它们各自产生的权重记作W_{21}、W_{22}、W_{23}、……W_{2N}。第M行晶体管在权重矩阵中第M行它们各自产生的权重记作W_{M1}、W_{M2}、W_{M3}、……W_{MN}。至此每个晶体管皆配置有贡献给参与网络计算的矩阵Ma_wei的一个相应权重值。传统晶体管自身并无与权重相关联的特征,尤其是闪存存储器与神经网络无直接相干性,所以如何将权重映射到晶体管是另一个需要解决的关键问题。

[0078] 参见图4,合计M行N列晶体管视为M×N的矩阵元素,矩阵Ma_wei中各个权重值的界定是实现神经网络NET的基础。图中展示每一行晶体管的源极电流流向该行晶体管公共的源线、图中展示每一列晶体管的漏极连到该列晶体管公共的位线,这种晶体管的布局将为如何给晶体管界定其权重带来可行性方案。例如每一行晶体管的源极电流汇聚特性对应于前文向量矩阵乘法的乘法及累加计算算法。

[0079] 参见图4,为了解释如何将权重映射到晶体管,先行介绍存储晶体管的电流和阈值以及施加电压之间的关系:栅极的阈值电压(threshold voltage)为V_{TH};存储晶体管与工艺制程相关的工艺参数K=μCW/L,W/L是晶体管的宽长比尺寸,对确定制备工艺及确定尺寸的存储晶体管其工艺参数K近乎为定值;栅极源极之间的电压V_{GS}和晶体管漏极与源极之间的电压V_{DS}可满足于$K(V_{GS}-V_{TH})V_{DS}=I$。晶体管流过的电流I在该晶体管工作在线性区或三极管区的时候,如果单个晶体管的电流I执行乘法运算以及一行晶体管它们各自的输出电流I执行累加,这等效于矩阵乘法的乘法及累加计算。

[0080] 参见图4,针对第一行晶体管(栅极耦合到G1):第一个晶体管的电流I1通过计算大约记作$K(V_{GS1}-V_{TH1})V_{DS1}=I1$,按照相同的电流公式,第二个晶体管的电流I2通过计算大约记作$K(V_{GS2}-V_{TH2})V_{DS2}=I2$、第N个晶体管电流$K(V_{GSN}-V_{THN})V_{DSN}=IN$。第一个晶体管栅源极之间的电压为$V_{GS1}$及漏源极之间的电压为$V_{DS1}$、阈值电压为$V_{TH1}$,第二个晶体管栅源极之间的电压为$V_{GS2}$及漏源极之间的电压为$V_{DS2}$、阈值电压为$V_{TH2}$,第N个晶体管栅源极之间的电压为$V_{GSN}$及漏源极之间的电压为$V_{DSN}$、阈值电压为$V_{THN}$。

[0081] 参见图4,针对第一行晶体管(栅极耦合到G1):第一个晶体管至第N个晶体管它们各自的源极电流皆流向该行晶体管公共的源线S1,源线S1的电流记作IS1。

[0082] $IS1=I1+I2+I3+\cdots IN$

[0083] 参见图4,针对第一行晶体管(栅极耦合到G1):第一个晶体管所在位置处如果定义了W_{11}与$K(V_{GS1}-V_{TH1})$成比例关系或相等、以及V1与V_{DS1}成比例关系或相等则流过前述第一个晶体管电流I1的可转折表达成$I1=W_{11}\times V1$。第二个晶体管所在位置处如果定义了W_{12}与$K(V_{GS2}-V_{TH2})$成比例关系或相等、以及V2与V_{DS2}成比例关系或相等则流过前述第二个晶体管电流I2的可转折表达成$I2=W_{12}\times V2$。第三个晶体管所在位置处如果定义了W_{13}与$K(V_{GS3}-V_{TH3})$成比例关系或相等、以及V3与V_{DS3}成比例关系或相等则流过前述第三个晶体管电流I3的可转折表达成$I3=W_{13}\times V3$。第N个晶体管所在位置处如果定义了W_{1N}与$K(V_{GSN}-V_{THN})$成比例关系或相等、以及VN与V_{DSN}成比例关系或相等则流过前述第N个晶体管电流IN的可转折表达成$IN=W_{1N}\times VN$。

[0084] 参见图4,针对第一行晶体管(栅极耦合到G1):源线S1的电流记作IS1经过式子的转换则可以表达成以下方式(还可以定义$Y_1=IS1$)。前提条件是第一行晶体管它们的源极电流汇聚特性,这对应于向量矩阵乘法的乘法及累加计算算法,因此,其属于闪存存储器能够替代传统CPU或GPU部分功能的主因素之一。

[0085] $IS1=Y_1=W_{11}*V1+W_{12}*V2+W_{13}*V3+\cdots W_{1N}*VN$

[0086] 参见图4,针对第二行晶体管(栅极耦合到G2):源线S2的电流记作IS2经过式子的转换则可以表达成以下方式(还可以定义$Y_2=IS2$)。

[0087] $IS2=Y_2=W_{21}*V1+W_{22}*V2+W_{23}*V3+\cdots W_{2N}*VN$

[0088] 参见图4,针对第M行晶体管(栅极耦合到GM):源线SM的电流记作ISM经过式子的转换则可以表达成以下方式(还可以定义$Y_M=ISM$)。

[0089] $ISM=Y_M=W_{M1}*V1+W_{M2}*V2+W_{M3}*V3+\cdots W_{MN}*VN$

[0090] 参见图4,前文定义X_1至X_N为神经网络NET的输入层、定义Y_1至Y_N对应就是神经网络NET的输出层。至此,闪存存储器的第一行晶体管完成了一次从输入层到输出层的向量运算Y_1,而且Y_1的结果可以通过感测第一行晶体管的源线S1的电流而一次性直接的计算出来,不必单独计算$W_{11}\times V1$或$W_{12}\times V2$或$W_{1N}\times VN$等,属于非冯诺依曼架构下的类计算,计算结果不是通过数值运算得出、是由电流感测得出。同理闪存存储器的第二行晶体管完成了一次从输入层到输出层的向量运算Y_2、闪存存储器的第三行晶体管完成了一次从输入层到输出层的向量运算Y_3、闪存存储器的第M行晶体管完成了一次从输入层到输出层的向量运算Y_M。在可选实施例中,电流IS1-ISM经过ADC器件例如和闪存存储器封装到一起的模数转换器之模拟量到数字量的转换处理,电流值各自对应的数字量适用于各类电子设备,ADC(Analog to Digital Converter)器件也还可以和闪存存储器在晶圆制造阶段而直接集成到相同的半

导体晶片上。

[0091] 参见图4,闪存存储器的晶体管阵列完成的一项任务是构建矩阵Ma_wei,而矩阵中的各个元素对应于神经网络NET的权重矩阵。

[0092] 参见图4,矩阵Ma_wei以神经网络NET输入层Lay1/隐藏层Lay2为例,当输入数据输入至输入层Lay1、在输入层Lay1和隐藏层Lay2之间存在着矩阵Ma_wei,而且输入数据折算成电压V1-VN以给到输入层Lay1,那么隐藏层Lay2所得到的输出等效成输入数据由权重矩阵Ma_wei执行矩阵运算所得到的运算结果。电压信号作输入的类型以及输出的数据为电流值信号,通过权重矩阵Ma_wei对输入的电压信号的作用从而影响到输出电流值的幅度,这也即通过NOR Flash存储器实现人工智能运算的大致过程。

[0093] 参见图4,闪存存储器除了用于上文AI人工智能计算外,存算技术也可用于类脑芯片等其他类计算方面的应用,代表着一类未来主流的大数据计算芯片架构等,同时亦可作为低能耗的超算中心的核心计算模块。譬如NOR FLASH存内计算可以在单芯片中支持到大约300M左右的深度学习权重参数,不需要额外的内存即可进行计算。

[0094] 参见图4,闪存存储器对于神经网络NET的计算与其他类型的通用计算相比具有下文描述的几个重要特性,也属于自身所体现出的优势。

[0095] 参见图4,第一个方面,基于闪存存储器的神经网络NET计算流程简单、并允许具有大量并行性。神经网络NET的计算不再采用通用计算惯用CPU之分支预测、推测执行等机制才能高效运行,该特性使得用于神经网络NET的闪存存储器硬件加速器可以尽可能的减少或脱离CPU或GPU的使用,从而进行低成本、高效率的计算。

[0096] 参见图4,第二个方面,神经网络NET执行的计算仅由几种单运算组成,例如矩阵乘法和向量运算、卷积核的应用以及其他的线性代数运算。使得神经网络NET加速器可以类似于GPU,以大量简单的计算单元来应对业界需要执行的复杂数据运算。

[0097] 参见图4,第三个方面,神经网络NET自身表现出了很大的稀疏性特征,其大量的参数和计算结果非常小,甚至趋近于零。利用稀疏性的特征可以使运算合理利用数据跳过乘以零的计算,从而提高整体计算速度,尤其是矩阵运算。

[0098] 参见图4,第四个方面,神经网络NET允许较低如低比特位的精度计算,例如对于大部分人工智能应用可以采用8比特位或4比特位(意味着小晶体管阵列)定点数进行量化而几乎不损失识别准确率,可以根据人工智能应用场景采用低精度定点数进行计算从而提高推理速度、降低功耗量。功率损耗是值得强调的一点,传统CPU/GPU在模型训练和推理阶段会产生极大的热量,大部分电力损失在处理器的散热量方面,本申请的类计算则完全避免了处理器针对单目乘法以及乘法结果的累加过程,以及减少处理器和存储器之间不必要的数据搬运环节,降低能耗至原来的1/10至1/100。

[0099] 参见图4,举例说明,智能语音模型的大小通常在数百K到数兆的大小,端侧的图像推理模型大小通常在数兆到数十兆之间,市场上的常规NOR FLASH的存储空间能够轻易胜任这种规模的模型量。所以即便是手持终端的NOR FLASH存内计算芯片也足以满足绝大部分AI场景的需求。在优化NOR FLASH结构的前提下,例如众多闪存存储器的串并联或级联连接,近乎可媲美GPU所能够达到的所有人工智能应用场景。

[0100] 参见图5,前文针对权重如W_{11}与晶体管参数如$K(V_{GS1}-V_{TH1})$成比例关系或相等的映射关系进行了阐释,在神经网络NET中,权重是需随时调整和更新的信息,所以值得详细阐

释的部分还包括如何对晶体管参数(或称权重)进行适应性修改的内容。尤其是神经网络包括反向传播,反向传播的核心即权重的更新和迭代,显然闪存存储器的本地计算有别于传统处理器那样的权重迭代。

[0101] 参见图5,晶体管的阈值电压(threshold voltage)为V_{TH},阈值电压如果被改变那么相当于$K(V_{GS}-V_{TH})$被改变,等效于改变与$K(V_{GS}-V_{TH})$成比例关系的权重值,所以更改阈值电压V_{TH}是可选的方式。闪存数据的写入和擦除是通过操控与控制栅极相关的浮栅内部的电荷的诸如电荷释放与遂穿来进行的。譬如在写入方式上,NOR Flash采用的是热电子注入方式,通过在写入的时候提高控制栅的电压,例如通过源极向浮置栅注入电荷则浮置栅极电荷量的改变即可更改$K(V_{GS}-V_{TH})$。再如在擦除方式上,NOR Flash采用的是隧道效应将浮栅中的电荷挪走的,例如施加至控制栅的电压采用负电压等。

[0102] 参见图5,以第M行(栅极耦合到字线GM)晶体管为例,在编程PG0的擦除动作中可以清除例如位线BN所确定的晶体管(权重为W_{MN})之浮栅电荷,所以它的阈值电压是降低的以及权重W_{MN}则是增加的。编程PG0的擦除典型的是在字线GM处施加负电压来使得晶体管的浮栅电荷产生量子隧道效应而进入晶体管所在的衬底。按照前文所言基于权重值修改的需求,相反在编程PG1的写入动作中可以注入例如位线BN所确定的晶体管(权重为W_{MN})的浮栅电荷,在这种情况之下,所以晶体管它的阈值电压是增加的及权重W_{MN}是降低的。再者在编程PG2的写入动作中继续注入例如位线BN所确定的晶体管(权重为W_{MN})的浮栅电荷,编程PG2相对于编程PG1属于微调,所以晶体管它的阈值电压是继续增加的及权重W_{MN}则是微细降低的。其中电荷注入编程PG1的电压较之微调的编程PG1的电压要高得多,编程PG2的小电压刻度是避免权重过调。实质上电荷注入编程可以采用不限于图中的更多注入次数,这样权重W_{MN}不仅仅可以实现满足于权重矩阵矩阵Ma_wei的多数值需求、还能满足权重值的高精度需求。擦除编程动作和电荷注入编程动作如图所示的执行多轮循环代表着权重的迭代。

[0103] 参见图4,在可选的实施例中,执行存内计算的方法包括:在闪存存储器中利用存储晶体管阵列作为神经网络NET的权重矩阵(weight matrix)之载体,设定神经网络的输入数据输送至存储晶体管阵列(transistor arrays)的系列位线B1-BN、以及还设定从存储晶体管阵列的系列源线S1-SM处撷取到该神经网络NET的输出数据。其中存储晶体管阵列当中任一晶体管的权重之大小利用其阈值电压进行调节,藉此来动态配置神经网络的权重矩阵。如由位线B1和源线S1所确定的一个晶体管T_{11}:晶体管T_{11}的权重用其阈值电压进行调节。位线B1和源线S2所确定的一个晶体管T_{21}:晶体管T_{21}的权重用其阈值电压进行调节。位线B1和源线S3所确定的一个晶体管T_{31}:晶体管T_{31}的权重用其阈值电压进行调节。位线B1和源线SM确定的一个晶体管T_{M1}:晶体管T_{M1}的权重用其阈值电压进行调节。所以任一晶体管的权重之大小利用其阈值电压进行调节,藉此来动态配置神经网络的权重矩阵(weight matrix)。N和M是正整数。再如位线BN和源线SM确定的晶体管T_{MN}:晶体管T_{MN}的权重用其阈值电压入浮栅电荷等进行调节。

[0104] 参见图4,在可选的实施例中,位线B1-BN和源线S1-SM所确定用于神经网络的存储晶体管阵列(transistor arrays)位于闪存存储器的同一物理块(block)上;以及神经网络的输入数据(例如向量Ma_in)、输出数据(例如向量Ma_out)各自所在的空间与存储晶体管阵列(transistor arrays)所在的存储空间分属不同物理块(block)。例如输入数据所在的存储空间与输出数据所在的存储空间属于相同的物理块、或者输入数据所在的存储空间与

输出数据所在的存储空间分别属于相同的物理块。但输入数据、输出数据各自所在的存储空间与存储晶体管阵列所在的存储空间分属不同物理块。这种存储空间的剥离主要是考虑到允许存储类空间和计算类空间具有不同的擦除操作和写入操作,相当于利用阈值电压进行权重调节(图5之PG0、PG1、PG2等)的权重配置操作与存储数据例如神经网络的输入输出数据等的数据读写操作是在不同的存储空间所执行的动作,权重配置操作与存储数据之读写操作彼此之间互不干扰影响。因为权重配置操作和存储数据之读写操作使用到的编程电压(或编程电流)要求使用的电压水准(或电流水准)不一。

[0105] 参见图4,在可选的实施例中,神经网络NET的原始的输入数据转换成电压值形式的输入量,以基于排序成向量的输入量输送至存储晶体管阵列的一系列位线。原始的输入数据例如包括图像或视频的像素值或灰度值等数据类型,原始输入数据转换成图示的电压值形式的输入量V1、V2、V3…VN。电压值V1输入至位线B1、电压值V2类似的需输入至位线B2、电压值V3输入至位线B3、电压值VN输入至位线BN。

[0106] 参见图4,输入以向量Ma_in=[B1=V1、B2=V2…BN=VN]表示。以基于排序成向量的输入量[V1、V2…VN]输送至阵列(transistor arrays)的系列位线B1-BN。

[0107] 参见图4,在可选的实施例中,神经网络NET的输出数据以存储晶体管阵列的输出电流表征,存储晶体管阵列的一系列源线的输出电流被视为存储晶体管阵列的排序成向量的输出量。例如存储晶体管阵列(transistor arrays)的系列源线S1-SM它们各自的输出电流分别记作IS1、IS2、…ISM。源线S1的输出电流IS1可经过ADC模数转换数字量以及类似的源线SM的输出电流ISM可经过ADC模数转换数字量。网络NET的输出数据以存储晶体管阵列的输出电流IS1、IS2、…ISM表征。

[0108] 参见图4,在可选的实施例中,存储晶体管阵列的系列源线S1-SM的输出电流被视为存储晶体管阵列的排序成向量Ma_out的输出量[IS1、IS2、…ISM]。也可以认为神经网络的输出以向量Ma_out=[S1=IS1、S2=IS2、…SM=ISM]表示。

[0109] 参见图5,在可选的实施例中,调节晶体管(如BN/SM所确定T_{MN})权重的方式包括改变浮栅注入电荷(如改变T_{MN}的浮栅电荷),藉由晶体管的浮栅注入电荷的改变而调节任一晶体管在存储晶体管阵列中所处位置(如BN/SM所确定的位置)处的权重大小例如调节T_{MN}的权重W_{MN}之大小。

[0110] 参见图5,在可选的实施例中,在神经网络NET的反向传播的阶段,通过重新设置存储晶体管阵列中各个位置处的权重大小而动态配置神经网络的权重矩阵。这里所言的阵列中的各个位置包括位线B1-BN和源线S1-SM相互交叉处的每一个位置,其实所言的各个位置也代表着权重矩阵中的各个元素的行列索引。例如权重矩阵中W_{MN}这个元素的行列索引是由位线BN/源线SM所确定或对应的位置。神经网络在反向传播算法中需反复迭代权重矩阵中的各个元素的大小值,直至符合要求。

[0111] 参见图4,在可选的实施例中,神经网络的输入数据输送至存储晶体管阵列的系列位线以及设定从存储晶体管阵列的系列源线处撷取到神经网络的输出数据:晶体管阵列中每行存储晶体管的源极相连并视为神经网络的一路输出数据的源线输出端、晶体管阵列中每列存储晶体管的漏极相连并视为神经网络的一路输入数据的位线输入端。另外存储晶体管阵列中每行存储晶体管的栅极相连而耦合到一路字线上。

[0112] 参见图3,在可选的实施例中,在存储晶体管阵列的位线B1-BN中的每一者处皆配

置有数模转换器DAC,数模转换器DAC用于将神经网络NET的输入数据转换成与位线的输入信号模式(例如位线要求电压信号作为输入量)相匹配的电压值。

[0113] 参见图3,在可选的实施例中,在存储晶体管阵列的源线S1-SM中的每一者处皆配置有模数转换器ADC,模数转换器ADC用于将源线处的输出电流格式的输出信号转换成与神经网络NET的数字输出量相适配的输出数据。例如源线SM处所配置的模数转换器用于将源线SM处的输出电流ISM格式的输出信号转换成与神经网络的数字输出量相适配的输出数据。电流之数字量(digital)有别于模拟量(analog)。源线S1处所配置的模数转换器用于将源线S1处的输出电流IS1格式的输出信号转换成与神经网络的数字输出量相适配的输出数据。其他的S2至SM-1亦如此。

[0114] 参见图4,在可选的实施例中,执行存内计算的装置包括:闪存存储器和输入单元以及输出单元等。关于装置中的闪存存储器部分,可由其存储晶体管阵列作为神经网络的权重矩阵之载体,设定神经网络的输入数据输送至存储晶体管阵列的一系列位线以及设定从存储晶体管阵列的一系列源线处撷取到该神经网络的输出数据。以及关于装置中的输入单元部分,可由其将神经网络的原始的输入数据转换成电压值形式的输入量,以基于排序成向量的输入量输送至存储晶体管阵列的一系列位线。关于装置的输出单元,用于将源线的输出电流格式的输出信号转换成与神经网络的数字输出量适配的输出数据,存储晶体管阵列的一系列源线的输出电流被视为存储晶体管阵列的排序成向量的输出量。输入单元和输出单元分别包括数模转换器和模数转换器。

[0115] 参见图4,在可选的实施例中,执行存内计算的存储系统包括:基于闪存存储器的存储芯片和内置于存储芯片的数模转换器及内置在存储芯片中的模数转换器。基于闪存存储器的存储芯片,其中在闪存存储器中利用存储晶体管阵列作为神经网络的权重矩阵之载体以及其中任一晶体管的权重之大小利用其阈值电压进行调节,藉此来动态配置神经网络的权重矩阵;设定神经网络的输入数据输送至存储晶体管阵列的一系列位线以及还设定从存储晶体管阵列的一系列源线处撷取到该神经网络的输出数据。关于内置在存储芯片中的数模转换器,其中在存储晶体管阵列的每个位线处皆配置有数模转换器,从而可以将神经网络的输入数据转换成与位线的输入信号模式相匹配的电压值。关于内置在存储芯片中的模数转换器,其中在存储晶体管阵列的每个源线处皆配置有模数转换器,以将源线处的输出电流格式的输出信号转换成与神经网络的数字输出量相适配的输出数据。

[0116] 参见图5,以往的电子设备中计算和存储是分开的,计算例如处理器和存储例如硬盘或内存等存储器各自分别专注于本职任务,计算和存储基本互不干涉对方。亦有厂商为了提高计算速度和计算能力而将处理器和存储器在物理位置上集成得比较近,例如在电路板的版图设计上它们之间相互靠拢、短线路、在封装阶段直接相互层叠布置。

[0117] 参见图5,算力提升一直是按照摩尔定律的节奏逐步推进,但主流计算范式始终遵循着冯诺依曼架构。存算一体作为一种新型的算力,业界主要关注狭义的存算一体也即利用存储介质进行计算但计算媒介方法尚未完全统一。存算一体作为新型算力,有望解决传统冯诺依曼架构下的编译墙、存储墙、功耗墙问题。存算一体在业界的共识,其基本已被确定为解决算力问题的关键技术之一。

[0118] 参见图5,本申请映射到存储晶体管阵列的神经网络之输入数据或输出数据可以是神经网络的某一层(输入层/隐藏层/输出层等)的数据,换而言之,允许存储晶体管阵列

仅仅只是神经网络的局部子网络结构。或者是,神经网络可以使用多个如图所示的存储晶体管阵列进行前后级的串并级联组合,由多个存储晶体管阵列组合成神经网络的局部子网络结构或组合成全部网络结构。

[0119] 参见图5,储存和计算一体将存储与计算有机融合,基于其巨大能效比提升潜力并有望成为算力时代的先进计算架构方案。但是存在着疑虑:如何应对存储器内部神经网络权重更新时所面临的权重偏移(例如W_{MN}发生的偏移)等亟待解决的疑虑,对存储器而言这属于难以察觉的一种隐匿性的难题,下文将会继续详细阐释。

[0120] 参见图6,已知可由位线B1-BN和源线S1-SM确定用于神经网络NET的存储晶体管阵列(transistor arrays),以该阵列为例来阐释权重偏移的问题。存储晶体管阵列的尺寸应和神经网络的实际应用场景相贴合,譬如神经网络往往需要对各种长宽大小的图像进行卷积计算以求得卷积结果。举例说明:即便是简单的28*28长宽的手写识别图像就可能需要提供约784个输入端口,显然这种情况下存储晶体管阵列的尺寸相当大。此时较为棘手的难题是字线的长度会随着存储晶体管阵列的增长的尺寸而变得非常之长,因为权重的调整需依赖于通过字线来对晶体管浮栅执行电荷写入或擦除的编程,过长的字线会在晶体管阵列上会造成极大的字线寄生电容C_P、导致权重编程发生误差。寄生电容至少包括耦合至字线的各个晶体管的栅源寄生电容和栅漏寄生电容及字线对衬底的电容。存储晶体管阵列是制备在半导体衬底上。编程误差也就意味着权重配置的偏离。值得注意的是在很多神经网络诸如智驾域视频感知中的卷积神经网络等场合,与晶体管阵列之输入端口数相关的输入图像分辨率惯用1628*1236、1920*1080、2048*1024、3840*2160等,无论图像长宽是否实施缩放都会伴随巨量的阵列尺寸、及催生着寄生电容C_P的复杂化,因为寄生电容之多寡和阵列尺寸之大小不是完全对应的关系。寄生电容既有并联电容又有串联电容例如晶体管处的本体电容以及还有耦合电容例如不同字线之间的寄生耦合电容,这给权重的迭代如反向传播算法带来了极大错误率,训练和推理都存在极大的不合理性。在存储器的常规性应用场景下并不会担心这种问题,因为存储器面临超大或海量数据存储应用会以分割的物理块Blokc或平面Plane或颗粒LUN来应对这种存储需求,但在神经网络场景下若将权重矩阵分割在不同物理区会导致权重元素编程条件不一致性趋于严重。

[0121] 参见图6,在可选的实施例中,针对存储晶体管阵列的任一字线上的任一晶体管在动态配置权重的阶段(动态配置权重是因为神经网络的各参数在持续迭代),该任一晶体管在执行任意一次权重调节之前,先将该任一字线上相邻前一次权重调节的其他晶体管所使用的编程电压的反向电压瞬态挂载一次到该任一字线上、再将为该任一晶体管所设定的编程电压施加到该任一字线上以执行该任意一次权重调节。目的之一在于抑制存储晶体管阵列的该任一字线处存在的寄生电容的寄生电荷所导致的权重更新迭代失败。寄生电荷通常受该任一字线上实施权重调节的其他晶体管之编程电压的残存影响。在高图像分辨率状况下假设神经网络预期迭代的训练次数epoch=5000,在如此之大的阵列尺寸和如此之多的迭代次数的叠加加持下任何一个字线的寄生电荷将迅速隐匿式失控,很难掌握字线的实际寄生电荷的真实值而且即便知悉了又很快会被下一轮迭代而刷新掉。

[0122] 参见图6,在可选的实施例中,关于如何较佳应对存储器内部神经网络权重更新时所面临的权重偏移等亟待解决的疑虑:基础方案是存储晶体管阵列中任一晶体管的权重之大小利用其阈值电压进行调节,藉此来动态配置神经网络的权重矩阵。进一步的优化是针

对存储晶体管阵列的任一字线上的任一晶体管在动态配置权重的阶段,该任一晶体管在执行任意一次权重调节之前,先将该任一字线上相邻前一次权重调节的其他晶体管所使用的编程电压的反向电压瞬态挂载一次到该任一字线上、再将为该任一晶体管所设定的编程电压施加到该任一字线上以执行该任意一次权重调节。

[0123] 参见图6,举例说明,晶体管阵列(transistor arrays)中晶体管T_{MN}之权重大小利用其阈值电压V_{THMN}来调节,晶体管T_{MN}是位线BN和源线SM确定的一个晶体管且它所在的字线是GM。设定正整数M=1、2、3……、正整数N=1、2、3……。藉此可动态配置神经网络NET的权重矩阵Ma_wei,V_{THMN}中V_{TH}是阈值而下标MN是索引并且其表示了任意某个晶体管的阈值电压及权重W_{MN}皆可通过索引位置来操纵调节。

[0124] 参见图6,举例说明,晶体管阵列(transistor arrays)之任一字线如GM上的任一晶体管如T_{MN}在动态配置对应的权重W_{MN}的阶段,该任一晶体管如T_{MN}在执行任意一次权重调节之前,譬如在执行次数为epoch=1455的权重调节之前,较佳的需先将任一字线即字线GM上相邻前一次权重调节的晶体管(假设GM上晶体管T_{M3}是T_{MN}执行本次权重调节如epoch=1455次调节的相邻前一次权重调节的那一个晶体管)所使用到的编程电压例如V_{PGM3}的反向电压瞬态挂载一次到该任一字线GM上。这之后,再将为该任一晶体管如GM上的晶体管T_{MN}所设定的编程电压如V_{PGMN}施加到例如字线GM上以执行该任意一次权重调节譬如执行次数为epoch=1455的权重W_{MN}的调节动作。抑制存储晶体管阵列的该任一字线GM处存在的寄生电容的寄生电荷导致的权重(如W_{MN})更新迭代失败并且值得强调的是:这种方案不适用于存储器内的数据存储应用,否则会较大概率的导致存储数据位的逻辑值发生意外翻转。注意字线GM的寄生电荷具有易失性特征但是这种易失性的失电速率相对于权重矩阵迭代的速度而言可能会低一些,这和存储晶体管阵列的晶体管部署密度和布线等构件的选材及晶体管之关键尺寸等相关。前述方案能尽量弥补和挽救权重的错误率,其和神经网络参数如权重的完善及训练相辅相成。前述方案的重要目标还在于:压缩该任一晶体管执行任意一次权重调节所设定的编程电压与该任一字线上相邻前一次权重调节的其他晶体管所使用的编程电压之间的台阶差,避免该任一晶体管执行任意一次权重调节所设定的编程电压耦合到该任一字线的相邻字线上。从而防止相邻字线上的部分晶体管的权重值发生非期望迭代的更改。

[0125] 参见图6,举例说明,前述方案极力压缩该任一晶体管如T_{MN}执行任意一次权重调节所设定的编程电压如V_{PGMN}与该任一字线如GM上相邻前一次权重调节的其他晶体管例如晶体管T_{M3}所使用的编程电压V_{PGM3}之间的台阶差(voltage step),避免任一晶体管例如晶体管T_{MN}执行任意一次(譬如执行次数为epoch=1455)权重调节所设定的编程电压例如电压V_{PGMN}耦合到该任一字线GM的相邻字线GM-1或GM+1等处。在存储晶体管阵列(transistor arrays)中如果电压V_{PGMN}耦合到GM-1或GM+1等位置,会不经意的将字线GM-1或GM+1上耦合的一些晶体管(如GM-1/GM+1与BN交叉处)的权重值实施非迭代式的主动更改并使得权重矩阵发生局部而非全局的隐匿修改。这种情况下神经网络无论是训练还是推理都会产生错误的结果。而本申请设计的前述方案则巧妙的化解了相邻字线上的部分晶体管的权重值发生非期望迭代的更改的问题。

[0126] 参见图6,基于前文,用训练样本对卷积神经网络进行训练,训练出的参数例如权重矩阵等对于测试样本具有良好的泛化性能。在抵御存储晶体管阵列中任一字线上的寄生

电容所造成的错误权重方面,具有相当的鲁棒性和挽回非意外偏移。属于一个表现性能良好的卷积神经网络或子网络。神经网络典型的适用于智能语音模型、端侧的图像推理模型和绝大部分其他AI人工智能场景。

[0127] 以上通过说明和附图,给出了具体实施方式的特定结构的典型实施例,上述发明提出了现有的较佳实施例,但这些内容并不作为局限。对本领域技术人员而言,阅读上述说明后各种变化和修正无疑将显而易见。因此,所附的权利要求书应看作是涵盖本发明的真实意图和范围的全部变化和修正。在权利要求书范围内任何和所有等价的范围与内容都应认为仍属本发明的意图和范围内。

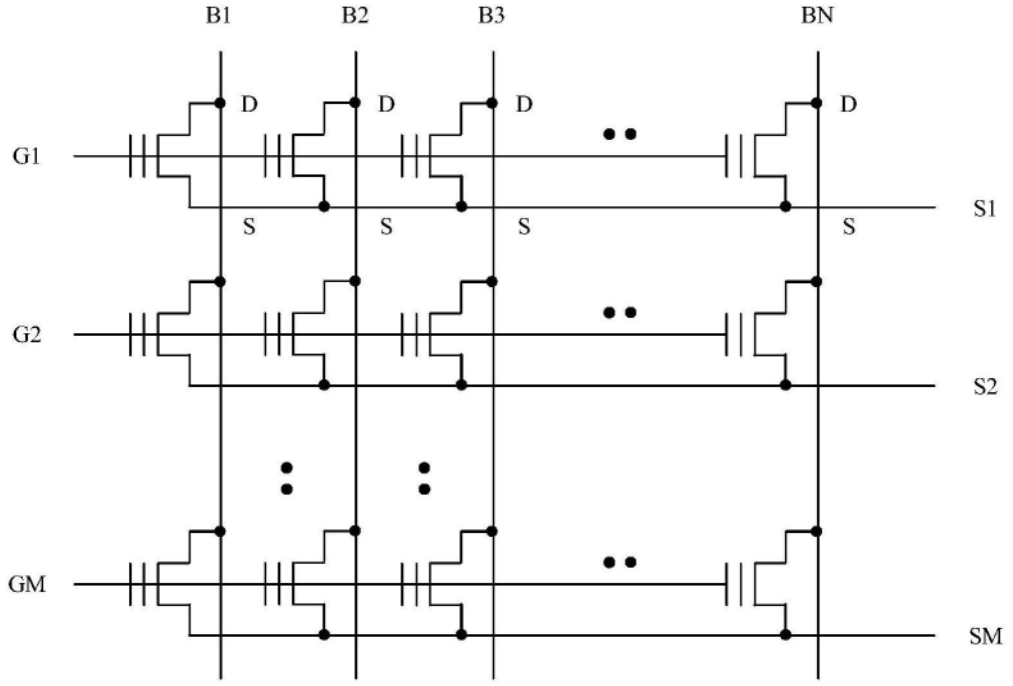

B1
B2
B3
BN
D
S
G1
G2
GM
S1
S2
SM

图1

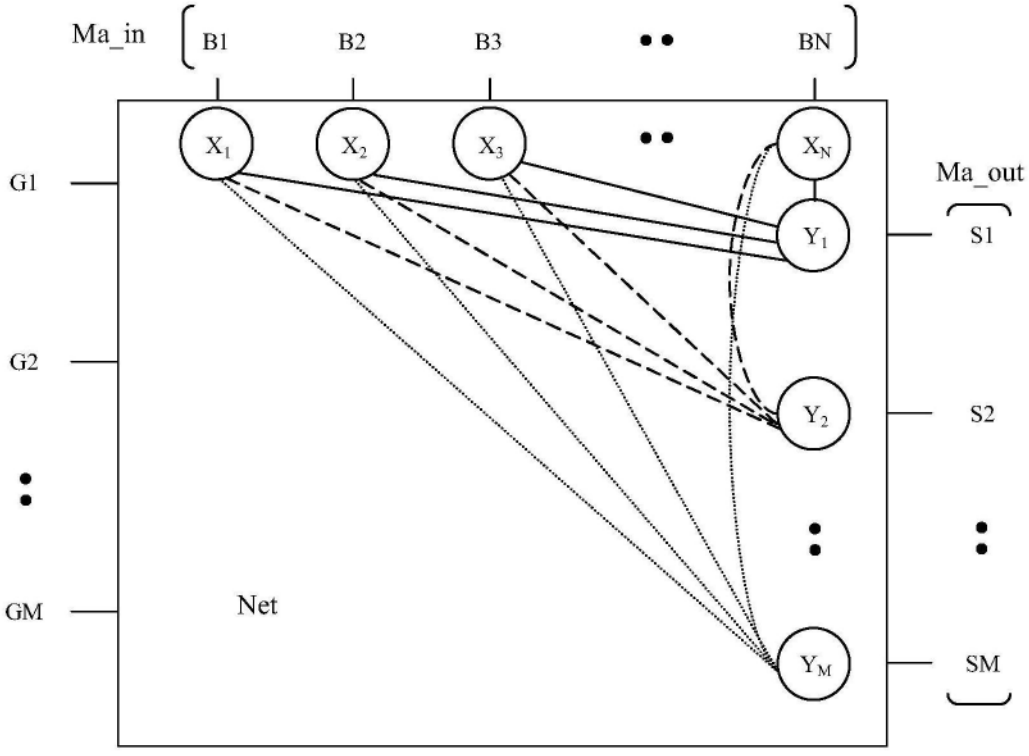

Ma_in
B1
B2
B3
BN
X1
X2
X3
XN
G1
G2
GM
Net
Y1
Y2
YM
Ma_out
S1
S2
SM

图2

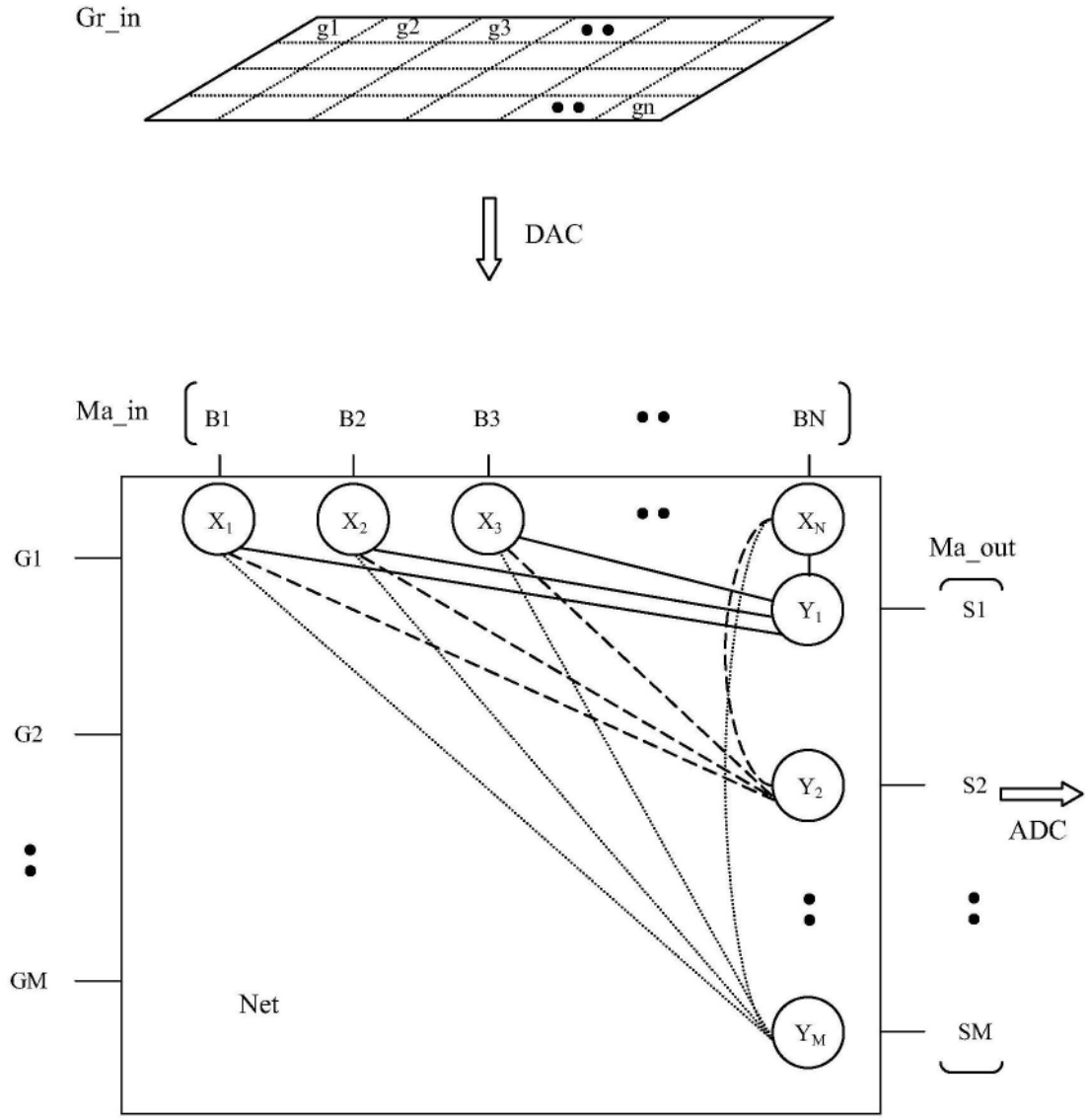

Gr_in
g1
g2
g3
gn
DAC
Ma_in
B1
B2
B3
BN
X1
X2
X3
XN
G1
G2
GM
Net
Y1
Y2
YM
Ma_out
S1
S2
ADC
SM

图3

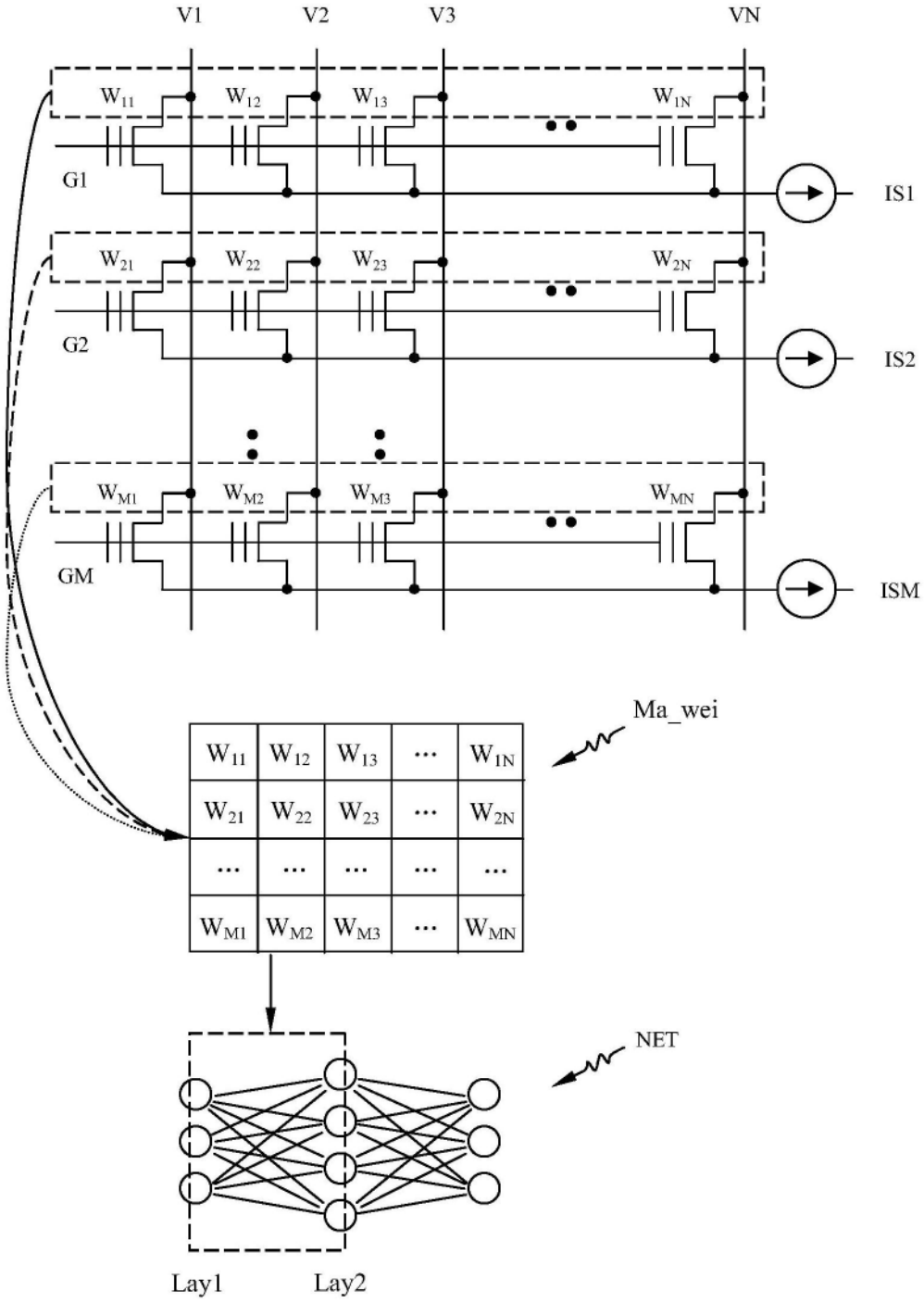

V1
V2
V3
VN
W11
W12
W13
W1N
G1
IS1
W21
W22
W23
W2N
G2
IS2
WM1
WM2
WM3
WMN
GM
ISM
Ma_wei
NET
Lay1
Lay2

图4

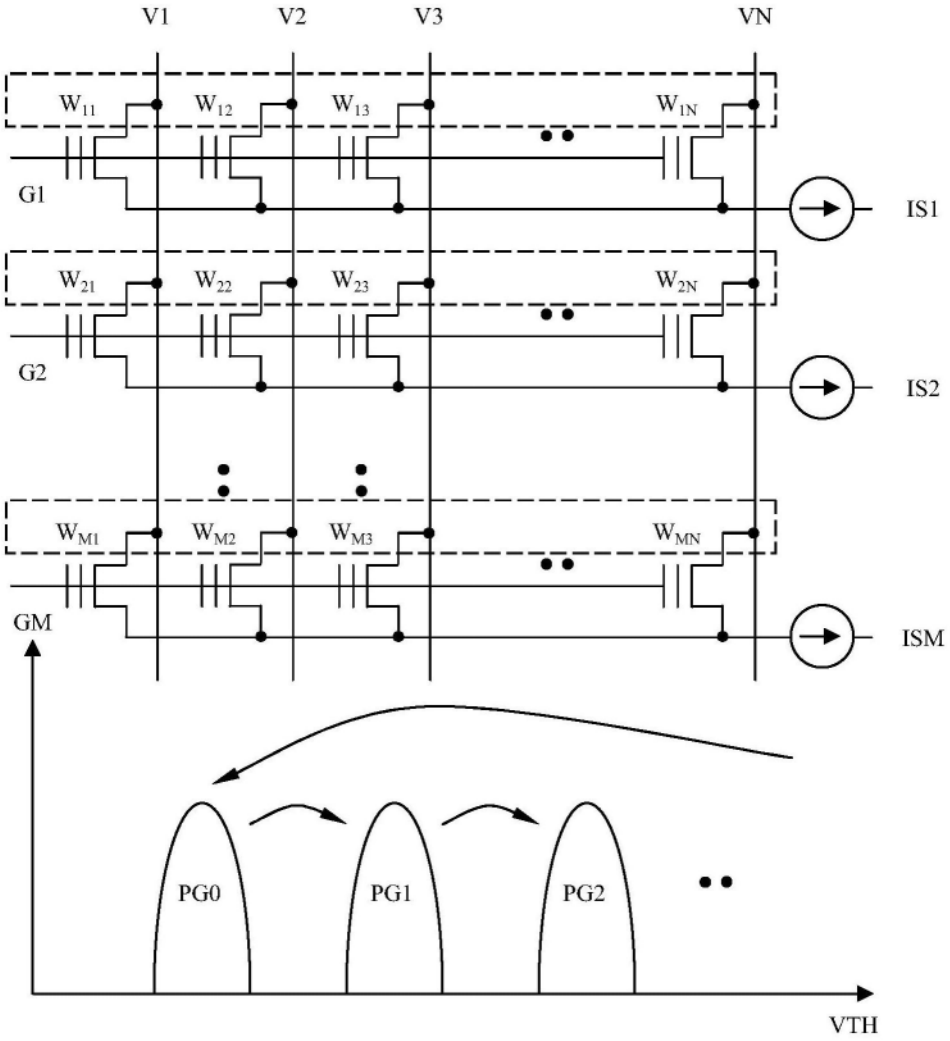

V1
V2
V3
VN
W11
W12
W13
W1N
G1
IS1
W21
W22
W23
W2N
G2
IS2
WM1
WM2
WM3
WMN
GM
ISM
PG0
PG1
PG2
VTH

图5

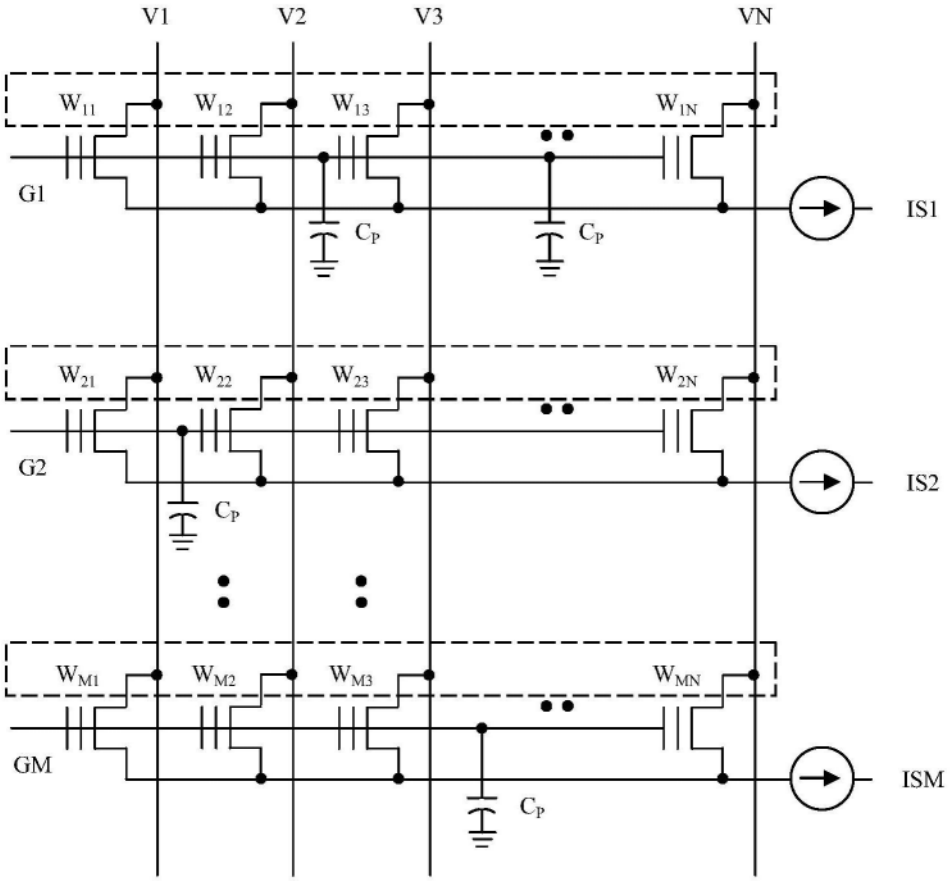

V1
V2
V3
VN
W11
W12
W13
W1N
G1
CP
CP
IS1
W21
W22
W23
W2N
G2
CP
IS2
WM1
WM2
WM3
WMN
GM
CP
ISM

图6