类脑神经网络与神经形态器件及其电路综述*

邓亚彬[a], 王志伟[a], 赵晨晖[a], 李 琳[a], 贺 珊[a], 李秋红[b], 帅建伟[c], 郭东辉[a†]
(厦门大学 a. 电子科学与技术学院国家示范性微电子学院; b. 萨本栋微纳米研究院; c. 物理科学与技术学院, 福建 厦门 361000)

摘 要: 为了系统地了解类脑神经网络电路,在对类脑神经网络进行简要介绍的基础之上,重点阐述两种类别的神经形态器件及功能,包括不同类型的浮栅管和不同工艺材料的忆阻器来模拟单个神经元和突触可塑性功能;然后,以神经形态器件为基础,分别介绍了基于浮栅管和忆阻器实现神经网络电路;最后总结当前神经形态器件及类脑神经网络芯片存在的问题,并对有关类脑计算研究方向进行了展望。

关键词: 类脑计算; 类脑神经网络; 神经形态器件; 神经网络电路
中图分类号: TP **文献标志码**: A **文章编号**: 1001-3695(2021)08-001-2241-10
doi: 10.19734/j.issn.1001-3695.2020.04.0349

Brain-like computing and neural morphologic devices with circuits

Deng Yabin[a], Wang Zhiwei[a], Zhao Chenhui[a], Li Lin[a], He Shan[a], Li Qiuhong[b], Shuai Jianwei[c], Guo Donghui[a†]
(*a. College of Electronic Science & Engineering, b. Pen-Tung Sah Institute of Micro-Nano Science & Technology, c. College of Physical Science & Technology, Xiamen University, Xiamen Fujian* 361000, *China*)

Abstract: In order to systematically understand the neural network circuits, this article began with a brief introduction of the principles of the brain-like neural network, as well as two types of neuromorphic devices and their functions, including different types of floating gates and memristors made by various materials to simulate individual neuron and synaptic plasticity. And then, it was based on the neuromorphic devices to introduce the circuit of the brain-like neural network implemented by floating gate and memristor respectively. Finally, it summarized the current problems of the neuromorphic devices as well as the brain-like neural network chip, and further made a prospect of the researching direction for the brain-like computing.

Key words: brain-like computing; brain-like neural network; neuromorphic devices; neural network circuit

类脑计算是目前人工智能研究领域的重点,其中类脑神经网络课题备受关注。类脑神经网络主要是借鉴人脑神经网络结构,应用神经形态器件模拟突触可塑性等类脑功能来实现智能化信息处理的人工神经网路。目前有关类脑神经网络的热点课题一般从两个不同领域开展相关研究。首先人们从模拟生物神经网络系统及结构出发,提出了以脉冲神经网络(spiking neural network,SNN)模型为主的研究方案,涉及拟态神经元及其突触的器件兑现方案与网络结构[1],其中神经元是以脉冲信号相互交流,具有更强的信息处理和容错能力[2,3]。其次,大部分有关人工智能的神经网络算法是基于冯·诺依曼结构的计算机编程实现的,但随着大数据信息处理复杂性不断增加,数据存储效率和计算处理能耗已经达到极限,人们开始意识到需要建立起基于存储—计算一体化类脑的信息处理方式,以实现成熟智能算法的神经网络[4]。这两方向的研究各有侧重,但都“相向而行”且相互借鉴,可预见的是类脑计算将会模拟神经元系统结构来支持人脑感知认知算法功能。类脑神经网络模型通常被描述为基于脉冲动力学驱动的神经元与具有存储—计算一体化功能的突触组成并可实现并行分布计算处理的互连结构网络模型。而模拟实现生物神经元和突触的器件常称为神经形态器件,其中模拟突触[4]的神经形态器件是神经网络中信号传递与调控的基本单位,其权重可以反映神经元之间的信号相互作用[5,6],它是模拟人脑学习与记忆的重要基础[7,8]。近年来,不仅有基于浮栅晶体管的突触设计[9],也有新的基于忆阻器的突触设计[10,11]。忆阻器具有电导随外加电压或电流的变化而变化的特性,并且在偏置截止后电导保持不变。此外,基于记忆电阻器的模拟神经元和突触在电路面积和功耗方面均优于数字电路[12]。随着神经形态器件的不断创新,也推动类脑计算以极低的硬件代价向高智能化发展。

1 类脑神经网络简介

类脑神经网络主要是模拟人脑突触、神经元以及连接方式来实现神经形态计算。从生物学层面讲,神经元细胞的主体被称为体细胞(soma),通过树突(dendrite)、突触(synapse)等实现信号传递与调整等功能[13]。因此在介绍神经形态器件和电路之前,首先要了解神经元模型与神经信号处理的基础(即突触可塑性[7])、类脑神经网络计算模型。

1.1 神经元模型

在生物神经元中,信号以脉冲的形式出现,这种神经冲动被称为动作电位或峰值[14],而确定神经编码(即最小的一组能够编码相关刺激参数的响应模式)是理解感知功能的前提[15,16]。最新的研究表明,时间编码更接近于生物信号编码方法[17],其中适用于时间编码最常见的为集成激发模型(leaky integrate-and-fire model, LIF)[18]和脉冲响应模型(spike response model,SRM)[19]两种模型。由于类脑神经元模型之间的关联性,对 LIF 模型的微分方程进行求解,其表达式可以看成 SRM 模型的求解,所以类脑神经元最直观的模型称为脉冲响应模型(SRM)[20],如图 1 所示。

由上一个神经元激励产生的脉冲信号 $i_n(t)$ 经过突触加权

收稿日期: 2020-04-21; **修回日期**: 2020-06-19 **基金项目**: 国家自然科学基金重点项目(61836010);厦门大学创新团队培育计划项目(20720190116)

作者简介: 邓亚彬(1991-),男,湖北武汉人,博士研究生,主要研究方向为人工智能、模拟集成电路;王志伟(1996-),男,安徽芜湖人,硕士,主要研究方向为人工智能、模拟集成电路;郭东辉(1967-),男(通信作者),福建莆田人,教授,博导,博士,主要研究方向为电路与系统(dhguo@xmu.edu.cn).

产生突触后脉冲信号(post-synapic potential,PSP) $P(t)$,其中突触分为兴奋性突触和抑制性突触,分别产生兴奋脉冲信号(excitatory PSP,EPSP) 和抑制脉冲信号(inhibitory PSP,IPSP) 。而 PSP 的脉冲响应 ε 函数形式可以近似地描述为[20,21]

$$\varepsilon(s) = \frac{s}{\tau}\exp\left(1-\frac{s}{\tau}\right)H(s) \tag{1}$$

其中: τ 为时间常数; $H(s)$ 为 heavyside 阶跃函数,当 $s>0$, $H(s)=1$,否则 $H(s)=0$。可以看出函数中上升和下降时间是相互关联的,但在实际应用中希望两者时间能够独立变化。其功能可以描述为: 神经元的内部电势 $I(t)$ 是所有 PSP 信号的总和,当 $I(t)$ 大于神经元内部阈值 V_{th} 时,神经元会发出一个脉冲激励信号 $i(t)$,此时 $I(t)$ 会复位到一个静止水平,在这段时间会有一段不应期,即输入任何 PSP 信号,神经元不再有响应。

1.2 突触脉冲时间依赖可塑性

基于突触的可塑性特点,生物大脑可以实现多学习规则[7] ,其中最重要的一种称为脉冲时间依赖可塑性(STDP) ,它是调节神经元间连接的经验依赖性变化机制。通常来说,是指刺激两个神经元时,两个刺激时间差值发生改变时能够影响突触后电流的增强或者减弱效果[22~26] 。因此在类脑神经网络电路设计中一般都需要对 STDP 规则进行模拟。权重的变化可以通过时间学习窗来控制。经典 STDP 时间学习窗可以表示为[27]

$$\Delta w = \begin{cases} \Delta w^{+} = A^{+}\mathrm{e}^{\left(\frac{-\Delta t}{\tau^{+}}\right)} & \Delta t \geqslant 0 \\ \Delta w^{-} = A^{+}\mathrm{e}^{\left(\frac{\Delta t}{\tau^{-}}\right)} & \Delta t < 0 \end{cases} \tag{2}$$

一般情况下学习曲线可表示为图 2,其中突触权重的变化 Δw 表示为关于突触后神经元刺激信号的产生时间 t_{post} 与突触前神经元刺激信号产生时间 t_{pre} 之间时间差的功能, $\Delta t = t_{post} - t_{pre}$ 表示前后两个刺激时间差, τ^{+} 和 τ^{-} 表示 STDP 学习窗口时间常数, A^{+} 和 A^{-} 分别表示突触增强和抑制的最大权值变化。因此,在经典基于双脉冲的 STDP 中,当突触前脉冲先于突触后脉冲时,突触权重增加突触后电流增加,反之突触权重减少电流将会抑制。

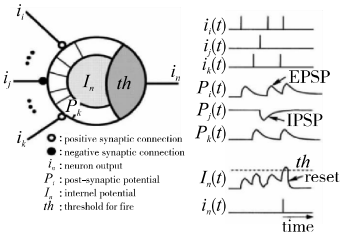


图1 脉冲响应模型

Fig.1 Spike response model

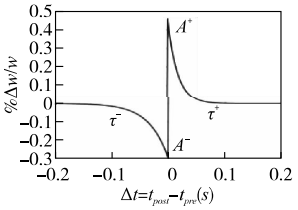


图2 STDP 学习曲线

Fig.2 STDP learning window

1.3 类脑神经网络计算模型

从上述生物神经元的结构和功能描述来看,单个神经元功能比较简单,但是通过突触互连的大规模神经网络却可以表现出复杂的类脑学习和认知功能如: 递归型神经网络[29] 、前馈型神经网络[30] 和自组织映射型网络[31] 等(图 3) ,都是对生物神经系统结构和功能的抽象模拟。尽管它们每个神经元的结构和功能相似,但各自构成的不同网络结构却能实现不同计算功能。其核心是模拟丰富的神经计算特性,而作为类脑神经网络的主体器件即神经元,它们在计算过程中也可表示为系统动力学行为过程的微分方程组,即各种神经元激活势能状态可表示为

$$\tau_i \dot{y}_i = -y_i + \sum_{j=1}^{n}\omega_{ji}\sigma(y_j - \Theta_j) + I_i(t) \tag{3}$$

其中: ω_{ji} 表示 j 和 i 神经元之间突触权重; $\sigma(*)$ 是阶跃或渐进型 sigmoid 函数; Θ_j 代表神经元阈值电压; τ_i 、I_i 分别代表神经元活跃时钟周期和神经元自身的刺激输入。同样,神经元之间的突触权重的学习过程也体现为离散的动力学过程,表示为

$$\omega_{ij}(t+1) = \omega_{ij} + \eta\frac{\partial C}{\partial \omega_{ij}} + \xi(t) \tag{4}$$

其中: $\xi(t)$ 表示随机扰动项; η 为学习率; $\omega_{ij}(t+1)$ 则是学习后突触权值; C 为目标函数可以理解为不同类脑神经网络功能的函数化表达。

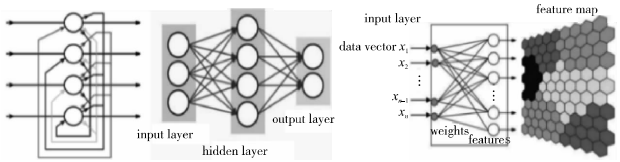


图3 三种基本神经网络拓扑结构

Fig.3 Three kinds of basic brain like neural network topologies

类脑神经网络计算可理解为: 不同目标函数表示不同神经网络的功能,其学习过程主要根据式(4) 对突触权重进行更新,并调整神经元的输出完成一次学习过程。通过迭代计算后,使神经元输出值收敛到稳定值完成整个学习过程。而类脑神经网络实现功能则是基于脉冲时间编码利用突触进行信息传递,并通过神经元动力学特性进行计算(式(3)) ,从而实现不同功能。

2 神经形态器件

目前对大脑的研究尚未完成,但类脑神经网络的实现可从模拟神经形态网络的结构和功能入手,主要在功能上模拟神经元与突触的信息处理方式。针对模拟生物基本信息处理单元功能的目标所研发的器件统称为神经形态器件,可分为突触器件、神经元器件两大类[4] 。目前多种模拟突触功能的 VLSI 电路被提出[32,33] ,神经网络向深度发展,突触器件的尺寸也越来越小[34] 。大量新型纳米尺度器件被研究与开发出来,这些器件可以近似模拟生物突触功能[35,36] ,在这些新型的基于纳米尺度的突触模拟器件中,忆阻器获得了广泛的研究。这是一种非易失性记忆元件,其状态/电阻可以通过施加足够强的电压脉冲而改变,是脉冲神经网络权重元件中最合适的候选器件。另一方面,忆阻器作为第四个基本电路元件被提出[37] ,它具有阻值连续可调、断电非易失、低功耗等特性,是模拟突触的最佳选择。

浮栅 MOS 管这种器件的结构和功能非常适用于人工神经网络,其具有多输入信号控制、阈值可变可控、电导可调节、制造工艺兼容、功耗低、可简化电路结构和节省芯片面积等特点,既能够作为神经元实现神经元功能[38] ,又能通过特定的结构实现突触功能,因此被广泛应用于类脑神经网络中。

2.1 浮栅管

神经元必须具有线性组合来叠加各个输入信号的加权和以及一个激活函数来限制神经元的输出。在往常的设计中,都需要通过复杂的模拟电路来实现部分的功能,在一定程度上已经成为了大规模集成神经网络发展的瓶颈。随着对新型浮栅管的进一步研究[39] ,浮栅管能很好地模拟突触和神经元的功能[40,41] 。

2.1.1 Neu 浮栅管

图 4 为多输入 Neu 浮栅管(也称为神经元 MOS[38]) 的结构。与 MOSFET 相比,它由多个栅极输入端组成,门极输入端与浮栅之间通过电容耦合,其中是浮栅电压, $C_i(i=1,2,\cdots,N)$ 是浮栅和第 i 个输入栅之间的电容, $V_i(i=1,2,\cdots,N)$ 是第 i 个输入栅上所加的电压,浮栅上的电压 φF 由式(5) 决定[42] 。

$$\varphi F = \frac{C_1V_1 + C_2V_2 + \cdots + C_NV_N}{C_{TOT}}, C_{TOT} = \sum_{i=0}^{N}C_i \tag{5}$$

当电压 φF 大于 Neu 阈值时,MOS 导通。即 Neu 导通条件为

$$\frac{C_1V_1 + C_2V_2 + \cdots + C_NV_N}{C_{TOT}} > V_{TH} \tag{6}$$

为了分析和观察神经元多输入加权求和以及阈值导通功能。假设 N 为 2, V_1 为信号输入端, V_2 作为神经元阈值调整端,由式(6) 可得

$$V_1 > \frac{C_{TOT}}{C_1}V_{TH} - \frac{C_2}{C_1}V_2 = 2V_{TH} - V_2 = V_{TH}^{*} \tag{7}$$

其中: $C_2 = C_1$; $C_{TOT} = C_1 + C_2$; V_{TH}为 MOS 管的固有阈值; V_{TH}^*为 Neu 浮栅管的等效阈值。当输入端和大于等效阈值时,MOS 管导通,等效阈值由阈值控制 V_2 的电压决定,因而实现阈值可编程功能。

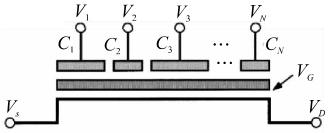


图4　Neu浮栅管结构

Fig.4　Structure of the functional MOS transistor

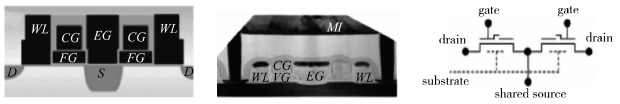
2.1.2　EFS 浮栅管

EFS 浮栅管是一种高度优化的浮栅单元(supercell) ,可以嵌入 CMOS 集成电路作为闪存器件。由于它是一种可调电导器件,当电路能对单个单元状态进行精确调整时,就能够在低功率亚阈值情况下对神经网络进行模拟矩阵乘法运算[43]。在类脑神经网络中,模拟电路在电路密度、速度和功耗能效方面远远超过具有相同功能的数字电路[44,45],由于其能够对每个单元进行单独调节,所以非常适合模拟突触器件[46]。

2017 年文献[43]基于 55 nm EFS NOR 闪存单元(图 5) ,制造并测试了一个 10×12 阵列模拟向量矩阵乘法器。改进后的阵列能够以低于 1% 的精度对每个单元进行高精度的单独模拟调谐,同时高度优化的单元状态能长期保持不变,且每个矩阵单元面积仅为 0.33 μm^2,通过栅极耦合附加外围单元,其精度约为 2%。同时,文献[46]基于该结构的模拟向量矩阵乘法器,并应用 EFS NOR 闪存单元设计出 BP 神经网络算法的硬件电路。该小组设计采用外围单元和矩阵单元的节能栅极耦合[43,47~49],如图 6(a) 在亚阈值模式下工作良好,存储单元的漏极电流 I_{DS}与栅极电压 V_G、阈值电压 V_T 几乎呈指数关系为[46]

$$I_{DS} \approx I_0 \exp\left\{\beta \frac{V_{GS} - V_T}{V_T}\right\} \tag{8}$$

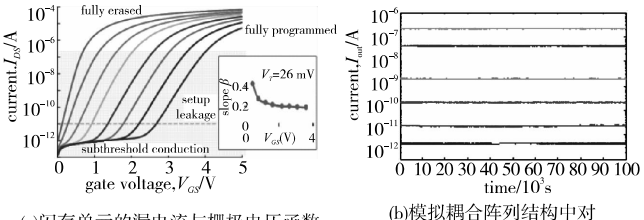
其中: I_0 为特征电流; β 为非理想因子,并通过测试表明单元对模拟量有几天的保存能力,输出电流波动非常小(图 6(b)) 。



(a)闪存单元　(b)闪存单元剖面　(c)闪存单元电路

图5　55 nm EFS NOR闪存单元

Fig.5　55 nm ESF NOR flash memory cells



(a)闪存单元的漏电流与栅极电压函数　(b)模拟耦合阵列结构中对存储器存储状态结果

图6　浮栅管I-V曲线与保持能力图

Fig.6　I-V curve and retention capacity diagram of floating MOS

2.2　忆阻器

2008 年 Strukov 等人[50]提出了忆阻器的数学模型,并应用纳米材料制造出器件,这个模型由高低阻态串联而成,其阻值随着通过的电荷量变化而改变[51]。之后,Strukov 等人[52]又提出了忆阻阈值开关模型,证实了忆阻器的边界和阈值效应,而忆阻器的阈值特性被认为是忆阻器件的一个重要而普遍的特性[53],其模型方程的表达式[50]为

$$v(t) = \left(R_{ON}\frac{w(t)}{D} + R_{OFF}\left(1 - \frac{w(t)}{D}\right)\right) i(t) \tag{9}$$

其中: D 为忆阻器的长度; ω 为忆阻器掺杂厚度(ω 会根据电荷的变化而变化) ; R_{ON}为导通低阻态; R_{OFF}为断开高阻态,并且分正负极。

通过给该忆阻器模型仿真,给忆阻器施加一个正弦电压, R_{on}和 R_{off}分别设置为 100 Ω 和 16 KΩ, $t_0 \equiv 2\pi/w_0 \equiv D^2/\mu_V \cdot v_0 =$ 10 ms, $i_0 \equiv v_0/R_0 = 10$ mA, $D = 10$ nm。其仿真结果如图 7 所示。因此在忆阻器的两端施加不同形状、时间等脉冲电压,能够模拟不同突触功能相应的神经刺激信号特点。

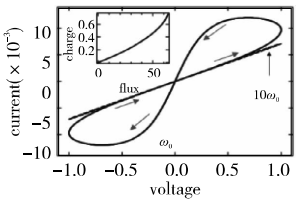
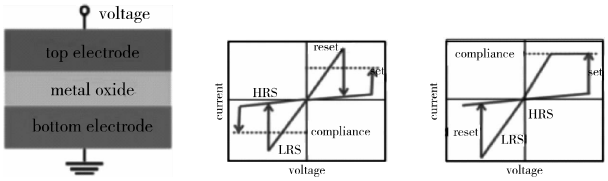


图7　忆阻器I-V曲线

Fig.7　Current variation at scaning voltage (-1 ~ 1V)

2.2.1　金属氧化物忆阻器

随着对忆阻器的进一步研究,以二元金属氧化物为代表的忆阻器即电阻开关随机存储器(RRAM) 被制造出来。由两个金属电极间夹着一层具有阻变性能的金属氧化物材料,如钙钛矿氧化物(PCMO) 、氧化镍(NiO) 、二氧化钛(TiO2) 等,金属氧化物材料的选择对 RRAM 的电流电压特性有很大影响。可以根据 RRAM 阻变器件的工作方式将其分为单极和双极两种。RRAM 阻变器件从高阻态(HRS) 转变到低阻态(LRS) 的过程称为置位(set) 过程,反之将阻变器件从低阻态回到高阻态的过程称为复位(reset) 过程,如图 8 所示[54]。



(a)金属氧化物RRAM结构　(b)单极工作方式I-V曲线　(c)双极工作方式I-V曲线

图8　金属氧化物RRAM的结构及工作方式I-V曲线

Fig.8　Structure and working mode of metal oxide RRAM I-V curve

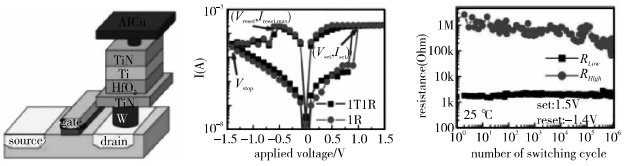
通常在应用 RRAM 时,会配置一个 NMOS 管与一个 RRAM 组成一个 1T1R 单元,Lee 等人[55]通过沉积后退火,用钛包覆层调整了 HfOx 的介电强度。在图 9(a) 中,给出了由 HfOx 薄膜和晶体管作为电流限制器(1T1R 配置比) 组成的典型 RRAM 单元的原理图。电阻开关器件的典型 I-V 曲线包括 1R 和 1T1R 结构,如图 9(b) 所示。由此产生的高速运行(<10 ns) 双极 HfOx RRAM 具有较大的开/关比(>100) ,可靠的开关次数 (>106 次) ,如图 9(c) 所示,且在高温下可靠性好,具有多比特存储和器件良率高的优点[54]。在同一工艺下使用 AlCu 和 Ta 作为 HfOx 的掩膜层,这些器件也显示稳定的双极电阻开关,由于掩膜层的氧捕获能力有限,导致其只有一个小的开/关比[55]。在超薄 HfOx(3 nm) 条件下,经过金属退火后 Ti/HfOx 堆栈呈现无成形特性[56]。最近文献[57]报道了以 Hf 为掩膜层的 10 nm TiN/Hf/HfOx/TiN RRAM。到目前为止,HfOx 是最成熟的 RRAM 材料之一。

2.2.2　石墨烯量子点(GQD) 忆阻器

在神经网络的构建中,突触的可控性是非常重要的,它有助于缩短训练周期、降低错误概率。有研究指出,记忆体器件的模拟电阻状态重复性差,需要大量的训练时间,导致工作效率低和能量浪费[58,59]。此外,它可能会增加神经网络的编程错误概率[2,60]。石墨烯量子点(GQD) 是一种边缘功能化的纳米尺度石墨烯碎片,由于量子约束和边缘效应,GQD 具有独特的光学和电子特性[61],所以被引入到忆阻器中。尽管记忆器件在可塑性和学习功能的展示方面取得了很大的成就,但是在记忆器件的可重复模拟电阻状态方面的研究进展甚微,因此石墨烯量子点忆阻器的研究是今后发展的一个重要方向。

2017 年,Wang 等人[62]基于标准半孔—导管工艺制造,将 GQD 嵌入在活性层(FeOx) 和顶电极(Pt) 之间研制出 GQD-Mem(石墨烯量子点忆阻器) ,并采用热重分析(TGA) 方法研

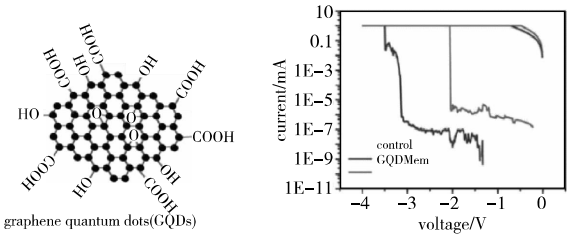
究 GQD 的热稳定性,得出 GQD 含有丰富的氧官能团,如羟基(－OH) 、羧基(－COOH) 、环氧化物(－O －) ,如图 10(a) 所示。采用电流为 1 mA 的电铸工艺初始化电阻开关,GQDMem 的形成电压约为 2.05 V(图 10(b)) ,远小于没有加入 GQD 的忆阻器(以下简称控制忆阻器) 形成电压约为 3.5 V,且在突变开关前 GQDMem 的隧穿电流高于控制忆阻电流。



(a)HfOx为典型的 RRAM 1T1R结构 (b)TiN/TiOx/HfOx/TiN器件在1R和1T1R结构下的双极电阻开关特性 (c)RRAM在500uS脉宽和106次开关频率下,高/低阻值变化率

图9 1T-1R单元结构及不同结构单元I-V曲线和电阻保持能力

Fig.9 1T-1R unit structure and I-V curves of different structural units and resistance retention capabilit



(a)含氧官能团的GQD (b)忆阻器与 GQDMem电铸工艺

图10 GQD原子结构及I-V曲线

Fig.10 GQD atomic structure and I-V curve

通过对 GQDMem 进行直流电压扫描,图 11(a) 显示了 GQDMem 和控制忆阻器 set/reset 的 I-V 曲线比较。GQDMem 的设置电压为 －0.7 V,复位电压为 1.0 V,绝对值均小于控制忆阻器的设置电压 －1.5 V 和复位电压 1.5 V,两者电压均降低 40% 左右。此外,GQDMem 的最大复位电流从 5 mA 降低到 2 mA。

工作电压和最大复位电流的降低表明,在记忆器件中引入 GQD 可以降低功耗。图 11(b) (c) 分别给出了 GQDMem 和控制忆阻器连续 50 个周期的 set/reset 重复性仿真。与控制忆阻器的集电压分布普遍形成对比的是 GQDMem 的 50 个周期的 I-V 曲线几乎完全重合。图 11(d) 显示 GQDMem 和控制忆阻的数据保持在程序状态下均表现出良好的非挥发性。

3 类脑神经网络电路实现

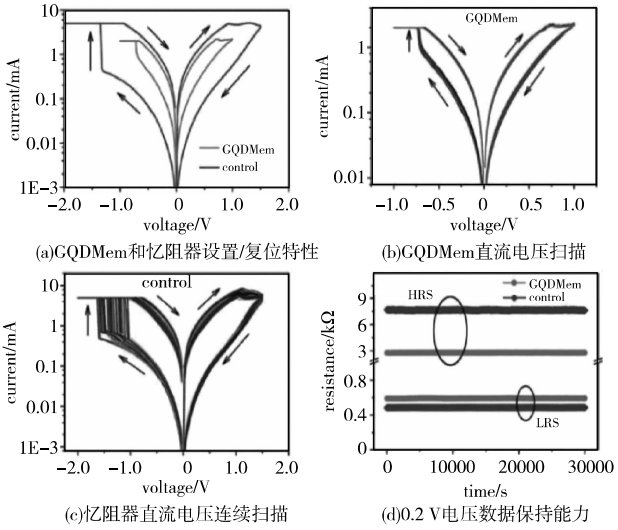
总体来说,根据类脑研究方向进行不同发展。一方面以大量人工突触及神经元连接而成的脉冲神经网络,通过时间编码模式对数据进行分类处理,并能够根据突触的可塑性实现人脑的学习规则。另一方面则是利用人工突触存储—计算一体化,通过现有的存储结构提高集成度并大大减小功耗,能够最大化并行实现当前成熟的神经网络模型。因此,本章以两种神经形态器件类别为基础,详细介绍脉冲神经网络电路和基于神经形态器件实现类脑神经网络计算的交叉阵列电路。

3.1 基于浮栅管的神经网络电路

浮栅管作为非挥发存储器件,能够与标准 CMOS 工艺兼容,并且能够通过改变浮栅中的电荷改变晶体管的电导,因此在模拟神经元和突触中有广泛的应用。另外基于浮栅结构的多输入器件(也称为神经元 MOS[38] ,Neu MOS) ,具有多端口输入和耦合电容加权形式进行求和,并且具有阈值可变、节省芯片面积等优点,能够很好地模拟神经元的功能。

3.1.1 神经元电路

神经元必须具有各个输入信号线性加权以及激活函数来限制神经元输出的功能。在以往的设计中,需要通过复杂的模拟电路来实现这部分的功能,在一定程度上已经成为发展大规模集成神经网络的瓶颈。因此应用多输入浮栅管作为单个神经元,能够集成大量的神经元实现特定的类脑功能。



(a)GQDMem和忆阻器设置/复位特性 (b)GQDMem直流电压扫描

(c)忆阻器直流电压连续扫描 (d)0.2 V电压数据保持能力

图11 GQDMem与忆阻器的I-V曲线及保持能力

Fig.11 I-V curves and retention of GQD and memristor

Morie 等人[63] 提出了一种多输入节点 Neu 浮栅管的 SRM 神经元电路,电路结构和参数如图 12 所示。在 Neu 浮栅管上构造了多个单电子电路。每个栅极输入端由一维(1 ~ D) 点阵列组成 $A_h(D_1,\cdots,D_n,D_C,D_n,\cdots,D_1)$,IN 为输入脉冲,V_p 为适当偏置,无脉冲时让电子高概率存在于中心点。因此输入脉冲都会影响单电子在数组 A_h 中的位置产生能量势垒,使其具有一定几率的跃迁。如图 13(a) 所示,假设输入信号 $V_p = V_{bg} = 0$ V,并且,基准电压 V_H 也设置为 0 V 或略小于 0 V。因此,由于单电子 eM 自身的电荷能,总能量随着 eM 在一维阵列 A_h 位置的变化曲线有两个峰。能量的最小值位于 D_c 和 D_1,D_c 能量势垒高度近似地由总电容和偏置电压决定,所以当隧道结节电容在 0.1 aF 左右时,能量差异大于室温热能量,可以减小温度对电路的影响。图 13(b) 显示当输入一个合适的宽度和幅值脉冲时,eM 能够快速地移动至 D_1,当脉冲信号停止后,eM 会通过热噪声辅助隧道效应使其跨越能量壁垒返回到中心点 D_c,根据这种特性能够使该电子电路产生类似神经元 PSP 功能。

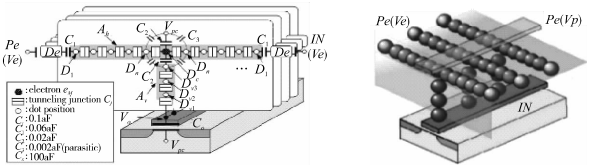


图12 基于Neu浮栅管和单电子电路的SRM神经元电路

Fig.12 A multinanodot Neu Floating-Gate MOSFET and single-electron for the SRM neuron

图 13(c) 显示该小组通过 100 个单电子电路并联到多输入浮栅管上,通过改变输入脉冲基准电压来影响电子跃迁概率,改变耦合到浮栅管的时间依赖电压 V_o,以实现神经元的 PSP 功能。

3.1.2 突触电路

突触在神经网络学习中起着重要的作用,根据突触前和突触后脉冲出现的时间差来改变突触强度的学习规则称为脉冲随时间变化的可塑性(STDP) 规则。浮栅管除了能够作为神经元实现脉冲响应外,同样能够实现突触的神经形态 VLSI,该晶体管可用于以非挥发性方式存储权重,并演示了生物学习规则,如长时程增强(LTP) 、长时程抑制(LTD) 和 STDP 等。

Liu 等人[64] 通过浮栅技术实现突触权值动态改变浮栅电荷数,以此来增大或减小权值更新的时间,即通过改变栅极的浮栅节点电压影响漏电流(即突触权重) ,并且能够通过电子隧穿或热离子注入[65,66] 改变浮栅电荷数,以此来增大或减小权值更新的时间常数,突触以及浮栅电压更新电路如图 14 所

示。图中(a) 为浮栅节点 f_g 根据(c) 输出信号进行更新; (b) 为电流型积分器作用于突触电路,偏置电压 V_τ 控制突触的时间常数[67]; (c) 用于产生控制脉冲的电路(vtunctrl 和 vinjctrl) ,分别在(a) 中打开隧穿调制和注入电流,并产生 P^* 和 M^* 信号作用于(d) 的 Soma 电路 vmem[64] 。为简化计算,引入额外变量 P 和 M 分别表示突触前后的活动[68] ,每次突触后神经元产生一个脉冲时,变量 M 都会有一定量地更新。相反,当一个突触每次接收到一个突触前的输入信号时,变量 P 也会相应更新。由式(2) 可知,A 是权值的更新函数,因此这些变量随时间衰减可表示为

$$-M(t) = \tau_- \frac{\mathrm{d}M}{\mathrm{d}t},\ -P(t) = \tau_+ \frac{\mathrm{d}P}{\mathrm{d}t} \tag{10}$$

其中: $P(t)$ 为突触后脉冲时采样; $M(t)$ 为突触前脉冲时采样,并且分别从突出权重增加或减少采样值。采样的 P 和 M 值调节 vmem 用于给单个类似于图 14(d) 中的 Soma 电路的电容充电。为了确定 vinjctrl 和 vtunctrl 脉冲产生时间,分别将 Soma 电路上的集成电压 $P(t)$ 、$M(t)$ 与对应的增强、减弱阈值 V_T 进行比较。当集成电压超过这个阈值时,激活 vinjctrl 或 vtunctrl 脉冲来更新 fg。实验仿真得到 $P^*(t)$ 与 $M^*(t)$ 的学习曲线采样图(图 15(a) ,其中输入脉冲序列频率(Pre) 为 50 Hz 和突触后频率(Post) 为 25 Hz) ,以及突触前后脉冲时间差与脉宽频率更新结果(图 15(b) ,突触前输入频率为 20 Hz,突触后输入频率为 20 Hz 的隧穿和注入脉冲频率,以及 τ_+ 和 τ_- 变化) 。以上都是以突触前、突触后两组脉冲时序实现 STDP(D-STDP[69,70]) ,而以往的研究表明[28,71] ,D-STDP 模型不能再现更高阶的脉冲模式(如三对和四对脉冲) 的实验结果,而且不能解释权值对脉冲的重复频率的依赖。

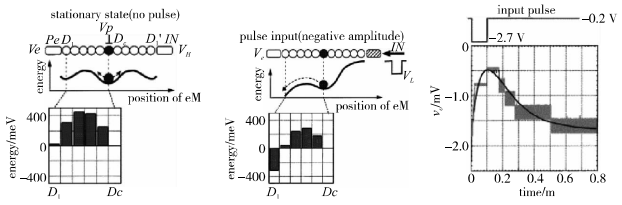


图13　不同电压的一维单电子能量图及输入脉冲与浮栅电压V_O的拟合α函数曲线(即PSP功能)

Fig.13　1D single-electron energy diagram with different voltages and a fitted-function curve is indicated by the solid line

Gopalakrishnan 等人[72] 在此基础上,用一个单浮栅晶体管(FG) 实现突触的微型化,该晶体管可以以非挥发性的方式存储权值,并根据三个脉冲时序来修改漏极电压,从而实现三脉冲 STDP(T-STDP) 学习规则。通过对式(2) 进行改进,T-STDP 规则可写成脉冲时间差函数[27,28] 。

$$\begin{aligned} \Delta\omega^+ &= \mathrm{e}^{\frac{-\Delta t}{\tau_+}}\left(A_2^+ + A_3^- \mathrm{e}^{\frac{-\Delta t_2}{\tau_y}}\right) \quad \text{if } t = t_{post} \\ \Delta\omega^- &= -\mathrm{e}^{\frac{\Delta t}{\tau_-}}\left(A_2^- + A_3^+ \mathrm{e}^{\frac{-\Delta t_3}{\tau_x}}\right) \quad \text{if } t = t_{pre} \end{aligned} \tag{11}$$

其中: A_2^+ 和 A_2^- 分别表示当存在 pre-post 对或 post-pre 对时权值变化的幅度; A_3^- 和 A_3^+ 分别表示电位和压降的三重项的振幅。$\Delta t_1 = t_{post}(n) - t_{pre}(n)$,$\Delta t_2 = t_{post}(n) - t_{pre}(n-1)$,$\Delta t_3 = t_{pre}(n) - t_{post}(n-1)$,τ_+ 、τ_- 、τ_x 、τ_y 为各个脉冲的时间常数。其表现形式如图 16(a) 所示,并且与 Pfister 等人[28] 的 T-STDP 时间规则进行对比。

图 16(b) 结构中新增加的模块(红色) 能使得三脉冲的权值从数学上简化分析。在新的结构中,每当突触前电位发生时,都会产生一个三角形的栅电压波形 V_g(图 17(a)) ,由于 MOS 晶体管亚阈值区域的栅电压与漏极电流呈指数关系,会产生一种类似于生物学的指数型兴奋性突触后电流。同样,每当突触后脉冲到达时,就会产生一个三角形电压 V_{tun} 和一个反向脉冲漏极电压波形 V_{d_inj}。在脉冲扩展(PlusExtender) 和多路复用器(Mux) 的作用下,对突触前脉冲发生时的全局三角形调制电压波形进行采样并生成电压 V_{tun_eff}(图 17(b)) ,图中红色部分为采样的隧穿 V_{tu_eff} 电压进行修正。因此,在新结构中,突触前电位后产生栅电压 V_g 和隧道电压 V_{tun_eff} 波形,在突触后电位后产生漏极电压 V_d 波形。然而,在之前的结构中,栅电压波形是在突触前电位后产生的,漏极电压波形和隧道电压波形都是在突触后电位时产生的。最后,通过测试得到单脉冲漏极电压和双脉冲漏极电压模式下的 STDP 仿真变化曲线如图 18 所示。

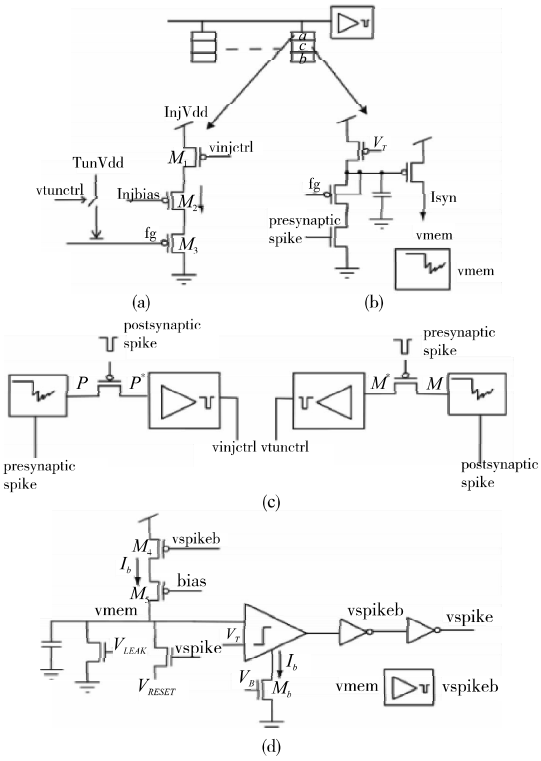


图14　突触电路结构

Fig.14　Synaptic circuit

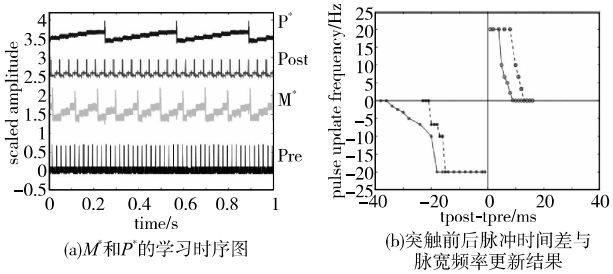


图15　学习规则实现

Fig.15　Learning rule implementation

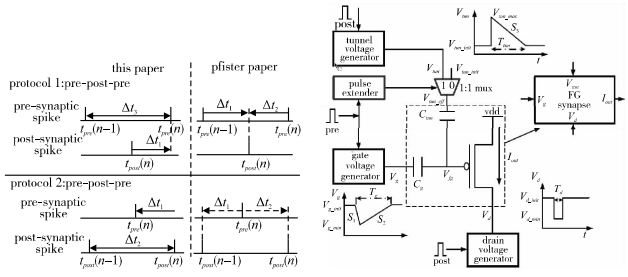


图16　不同STDP规则脉冲对比及其电路

Fig.16　Comparison of different STDP regular pulses and their circuits

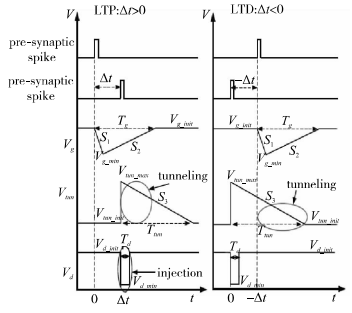
3.1.3　神经网络电路

为了实现人工神经网络算法,文献[46]利用商用的 180 nmESF 浮栅管搭建了三层神经网络形态电路,其中包含 101 780 个浮栅单元,并且总面积小于 1 mm^2,分类准确度实测中达到 94.7%,而模式的分类需要 1 μs 级的时间和 20 nJ 级的功耗。

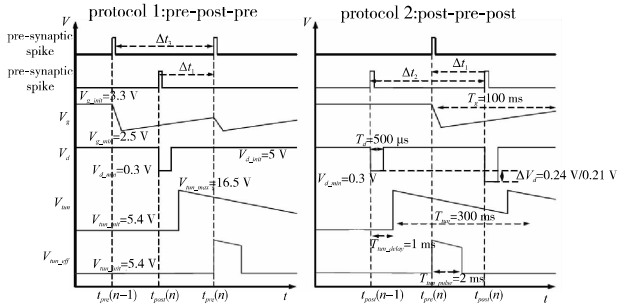
该小组通过最简单的三层神经网络模型,由一个三层感知网络与 784 个二进制输入 b_i 组成,并通过 MNIST 基准集(图 19(a))测试。例如,28×28 黑白像素的输入图像,64 个隐层神经元激活函数,10 个输出神经元(图 19(b))。该网络是通过对输入信号进行以下顺序转换来实现模式推理。其中 h_j 和 $f_j(j=1,2,\cdots,64)$,分别为隐层神经元的输入和输出信号,$c_k(k=1,2,\cdots,10)$ 为系统输出信号,提供输入模式的类,其函数可表达为

$$h_j=\sum_{i=1}^{784}\omega_{ji}^{(1)}b_i+\omega_{j,785}^{(1)},c_k=\sum_{j=1}^{64}\omega_{ki}^{(2)}f(h_j)+\omega_{k,65}^{(2)}f_{max}$$

$$f(h)\equiv f_{max}\times\begin{cases}\tanh(h) & h\geqslant 0\\ 0 & h<0\end{cases} \tag{12}$$



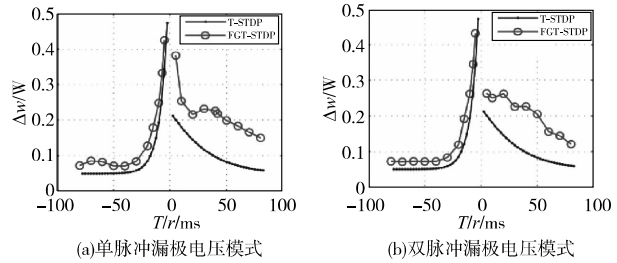
(a)D-STDP模型的LTP与LTD规则



(b)T-STDP模型时序图及电压变化曲线

图17 FG突触时序图及电压参数

Fig.17 Timing diagram and voltage specifications of FG synapse



(a)单脉冲漏极电压模式 (b)双脉冲漏极电压模式

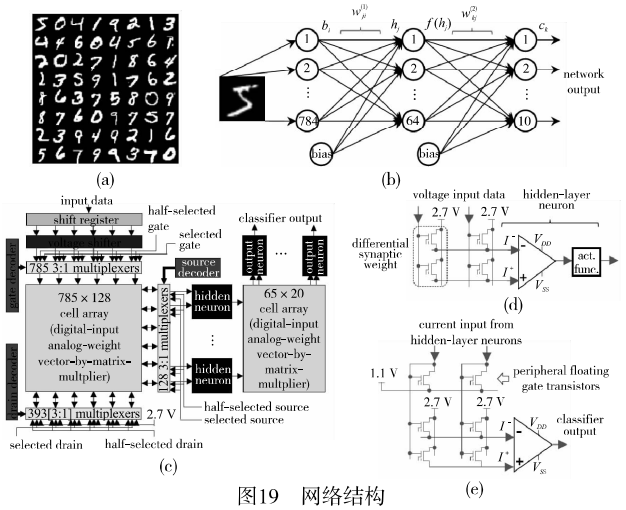
图18 不同脉冲模式突触STDP曲线

Fig.18 The synaptic STDP curve in different drain waveform

其中:$\boldsymbol{\omega}^{(1)}$ 和 $\boldsymbol{\omega}^{(2)}$ 是突触权重可调的两个矩阵,作为相邻的网络层的耦合特征。这些权值由具有可调权值两个横矩阵的浮栅门单元提供(图 19(c))。每个神经元从一个偏置节点获得一个额外的输入,该偏置节点具有基于类似细胞的可调权重(图 19(b))。第一横阵列中的矩阵乘向量混合信号是通过将输入电压(黑色像素为 4.2 V,白色像素为 0 V)直接施加到矩阵单元晶体管的栅极上实现的,其源电压(1.65 V)和漏电压(2.7 V)为固定电压(见图 17(d))。目的是使当前晶体管源漏电流(即矩阵的第 i 列和 j 行的交叉点)不依赖于任何其他单元的状态,其大小等于预设存储单元模拟权重 $\boldsymbol{\omega}^{(1)}$ 与二进制输入电压 b_i 的积。每一行的晶体管的源连接到一个单一的线(有外部固定电压),矩阵 j 行输出电流为所有 i 列 $\omega_{ji}^{(1)}b_i$ 的积求和,从而实现式(12)描述的向量—矩阵乘法,并且使用差分方案也能减小随机漂移。图 19(e)采用浮栅电容耦合方法对第二阵列进行模拟矩阵向量乘加计算,突触门阵列由额外与矩阵单元晶体管物理特性相似的"外围"单元行补充,因此有相同的亚阈值斜率 β。每列"外围"单元的栅电极与所有该列的晶体管栅极连接使其 V_{GS} 相等,通过式(8)可得到权重 ω 为

$$\omega_{kj}^{(2)}\equiv\frac{I_{kj}}{I_j}=\exp\left\{\beta\frac{(V_t)_j-(V_t)_{kj}}{V_T}\right\} \tag{13}$$

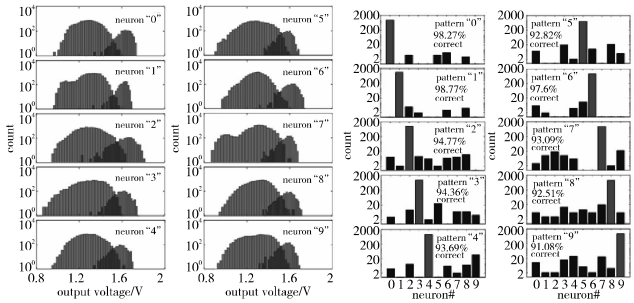
其中:I_{kj} 表示第 k 行 j 列的漏电流;I_j 为 j 列的总电流。同理,$(V_t)_j$ 与 $(V_t)_{kj}$ 分别表示 j 列的阈值电压与 k 行 j 列浮栅单元的阈值电压。通过实验并对输出 10 个神经元进行测试,得到每个神经元的正确率(图 20)。



(a) (b) (c) (d) (e)

图19 网络结构

Fig.19 Network architecture



(a)各神经元输出电压的直方图 (b)每个类的所有测试模式的最大输出电压的直方图

图20 实验10 000个MNIST测试集模式分类

Fig.20 Experimental results for the classification of all 10 000 MNIST test set patterns

3.2 基于忆阻器的神经网络电路

为了在神经网络硬件中实现突触的高密度集成,人们考虑各种纳米尺度的器件。这种设备除了表现出与脉冲时间相关的可塑性(STDP)外,还需要高度可伸缩及较大的耐受性,并且在状态间转换时需要较低的能量。而忆阻器的特性能够很好地满足以上条件,并且在类脑神经网络中得到广泛的应用。

3.2.1 STDP 突触电路

Shukla 等人[73]应用多个并行 PCMO-RRAM 作为突触模型,对 SNN 进行 STDP 规则训练,并进行演示不同数量的 RRAM 作为单个突触对学习效率的影响。该小组首先对 STDP 的学习函数(式(2))使用了一个乘法指数权重上升[74,75]函数包含指数相关项表示为

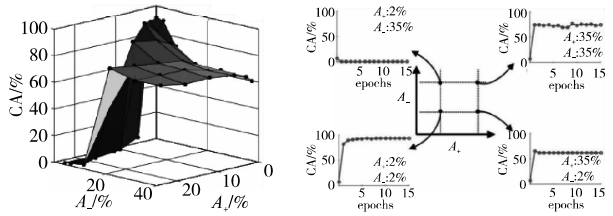
$$\Delta G(\Delta t,G)=\begin{cases}A_+\cdot S_+(G)\cdot e^{\frac{-\Delta t}{\tau^+}} & \Delta t\geqslant 0\\ A_-\cdot S_-(G)\cdot e^{\frac{\Delta t}{\tau^-}} & \Delta t<0\end{cases} \tag{14}$$

其中:$A_\pm$ 为比例因子(也称为学习速率);$S(G)$ 为生物饱和效应的模型[6,68,69];ΔG 为权值变化并且规定最大的 $G_{max}=1$,$G_{min}=0$,$A_{max}=1$;$S(G)$ 可以表示为

$$S_+(G)=(1-G)^{1.5}\&S_-(G)=G^{1.5} \tag{15}$$

图 21(a)显示了不同值的学习速率 $A_\pm$ 对分类准确度的影响,可以看到值越小准确率越高,图 21(b)则具体显示四个不

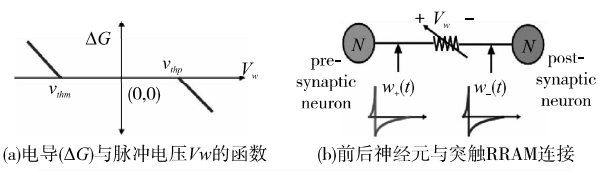
同增强抑制的突触权值变化率 $A_\pm$ 对分类精度的影响,从图中可以看出 $A_\pm$ 的变化率小于 2% 时,分类的准确度才能达到要求。图 22(a) 为单个突触的连接电路分为两个阶段。首先进入写阶段,即突触前后电位同时在 RRAM 两端输入(图 22(b)) 前后两个神经元信号,RRAM 会将两个信号根据时间 t 进行求和(图 22(c)) ,$\tau_\pm$ 为衰变常数,A_p、A_n 为增强、减弱幅值,$T_{p,n}$ 为脉冲信号增强或减弱的时长; 当超过 RRAM 的两个阈值 V_{thp} 或者 V_{thn} 时,改变 RRAM 的阻值状态(图 22(d) 红色圆圈部分) 。最后进入读阶段,即 RRAM 阻值状态改变后,输出脉冲信号。



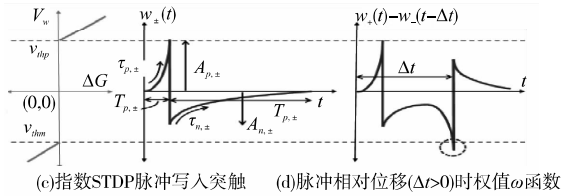
(a)分类准确度与学习速率$A\pm$%函数　(b)不同$A\pm$值的分类准确度

图21　学习速率与分类准确度关系

Fig.21　Diagram of learning rate and classification accuracy



(a)电导(ΔG)与脉冲电压V_W的函数　(b)前后神经元与突触RRAM连接



(c)指数STDP脉冲写入突触　(d)脉冲相对位移(Δt>0)时权值ω函数

图22　STDP突触电路及权值变化曲线

Fig.22　STDP synaptic circuits and weight change curves

该小组为了降低突触权值 A 的变化率,利用多个 RRAM 组成一个突触结构,读写分开对权值变化率进行更新(图 23(a) (b)) ,图 23(c) 显示其通过 MOSFET 开关进行逻辑控制 S_i 和 S_j 将写脉冲应用于相应的行(列) 。但在读取阶段所有开关都导通。因此,对于一个突触内的所有 RRAM,突触前/后端口接受相同的读取脉冲。降低突触权值 G^{ij} 的表达式可改写为

$$G^{ij}\rightarrow\frac{G_{max}}{N}\sum_{k=1}^{N}G_{RRAM}^{k,ij} \tag{16}$$

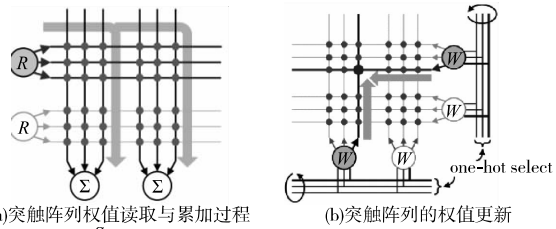
其中: N 为多个 PCMO-RRAM 共同作为一个突触。对于每一个峰值瞬间,基于 STDP 规则,权重 G 更新可表示为

$$G_{RRAM}^{K,ij}=G_{RRAM}^{K,ij}+\Delta G(\Delta t^{ij},G_{RRAM}^{K,ij}) \tag{17}$$

通过实验,对 N 取 2,16 和 100 进行训练,最后测试分类正确率(图 23(d)) ,可以看出 RRAM 越多的突触学习越稳定,正确率越高。

3.2.2　监督学习电路

Wang 等人[76] 利用 RRAM 开关突触设计了一种基于时空模式的神经形态电路。通过具有电阻开关突触的多神经网络的监督学习,实现时空模式下脉冲神经网络模拟人脑对声音的方位进行检测。其中脉冲信息依赖于一个简单的空间编码,即神经元同步的脉成一个只有空间的模式[77~79] 。该小组设计出基于 RRAM 突触的脉冲神经网络,能够进行时空模式编码,极大地提高了神经形态硬件处理信息效率。为了实现对突触权值的时变控制,将 RRAM 与图 24(a) 突触电路中的场效应晶体管(FET) 串联起来。单晶体管/单电阻(1T1R) 突触由 PRE 通过 FET 栅端子控制,并通过 FET 源端将突触电流 I_{syn} 传递到 POST。预处理电路中的轴突终端将尖峰形状为指数衰减脉冲 V_{axon} 来控制栅端产生漏电流。因此,脉冲前电压 V_{axon} 和脉冲后 VTE 之间的时间延迟导致 RRAM 电路的时间依赖可塑性即 STDP,如图 24(b) (c) 所示,这构成了 RRAM 突触处理信息的基础[80] 。

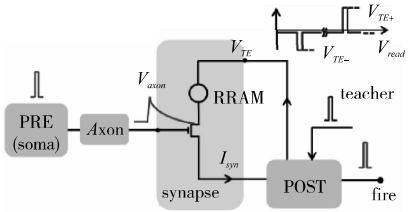


(a)突触阵列权值读取与累加过程　(b)突触阵列的权值更新

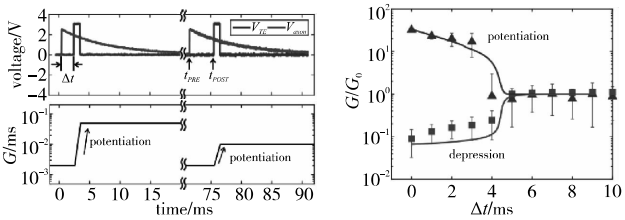
(c)突触阵列读写的开关电路　(d)相同结构中不同N值(2,16,100)与分类正确率关系

图23　突触网络电路及分类正确率

Fig.23　Synaptic network circuits and classification accuracy



(a)连接前轴突和后轴突的1T1R突触



(b)V_{te}与V_{axon}时间差,对突触电导增强的影响　(c)电路(a)的STDP

图24　神经元与突触电路

Fig.24　Neuron and synaptic circuit

为了识别特定的时空模式,必须对网络进行指示,使突触权值与真实序列呈现正相关关系。为此,权重由 Widrow-Hoff (WH) 学习规则[81] 更新。采集到序列中的每个脉冲后,V_{int} 达到阈值后,POST 神经元将其输出信号与 teacher 信号进行比较,如图 24(c) 所示。如果 POST 与 teacher 脉冲都存在,则正确检测到真实模式(true fire) ,因此不需要进一步更新权重。如果只存在后脉冲信号,则为错误的模式(false fire) ,通过施加振幅 $VTE<0$ 的负更新脉冲(图 25(a)) 来减小突触的权重,从而导致复位过渡。最后,如果只存在 teacher 信号,则为错误沉默模式(false silence) ,因此需要加强突触权重,以实现与突触序列的正相关。该小组采用第 1 层有 16 个 PREs 神经网络,通过 16 个 RRAM 突触与 1 个 POST 的第 2 层完全连接(图 25(a)) 。通过 Arduino Due 单片机作为监控器,生成输入和 teacher 信号图(25(b)) 用于监督训练和随后的模式识别测试。在每个训练周期中,16 个 PRE 接受以四个突触 PREs 为一组(cycle) 的时空编码模式,其中 1-4-9-16 为正确序列,每组对比后标记为 true、false、false silence 模式(图 25(c)) 。在每个模式中,用 1 ms 的时间间隔来分隔峰值,而暂停 50 ms 用来区分子序列模式,并允许恢复轴突电位 V_{axon} 和内部电位 V_{int} 的剩余状态。通过多次训练如图 25(d) 所示,可以发现#1、#4、#9、#16 突触权值在 false silence 模式下变大,在 false 模式中变小。当训练完成后可以发现#1、#4、#9、#16 突触处于低阻态,并且电导为增加趋势,而由于突触时间依赖增强,其他的突触处于高阻态。

因此该 SNN 中 POST 神经元的内部信号 V_{int} 提供了一个通

用测试序列和真实序列之间相似性的差异图。该相似性的差异可以应用到任何类型的序列，并且基于时空 SNN 与生物神经系统的相似性可以进一步说明，它能够检测精确的脉冲时间间隔，这模拟了大脑对声音位置的检测[82]。如图 26(a) 所示，当人耳接收到声音的时间顺序为 $ITD = t_L - t_R$，然后通过 2 × 2 的 RRAM 突触进行信息处理(图 26(b)) ，突触权重矩阵是对角的，因此一个触发后响应为左/右模式，而另一个触发响应相反的序列以此增加准确性。当音频信号触发响应的前神经元脉冲信号(图 26(c)) 引起左右轴突 V_{axon} 时间差异(图 26(d)) 并将时空模式分配给网络，使两个训练后的 POST 神经元内部二次脉冲电位之间的差值 ΔV_{int}(图 26(e)) 提供了声音方位的指示，通过测量与计算 ITD 与 ΔV_{int} 关系得到声音的方向角 θ。因此能够通过脉冲神经元电路实现类似人脑对声音方位的辨识。

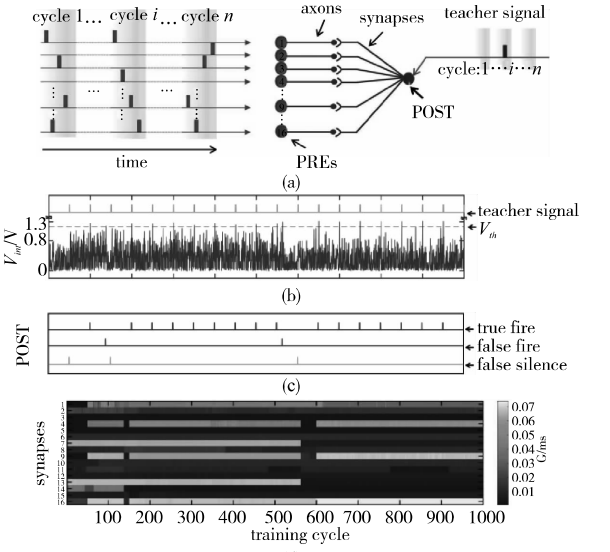


图25 时空模式的实验学习

Fig.25 Experimental learning in spatio temporal mode

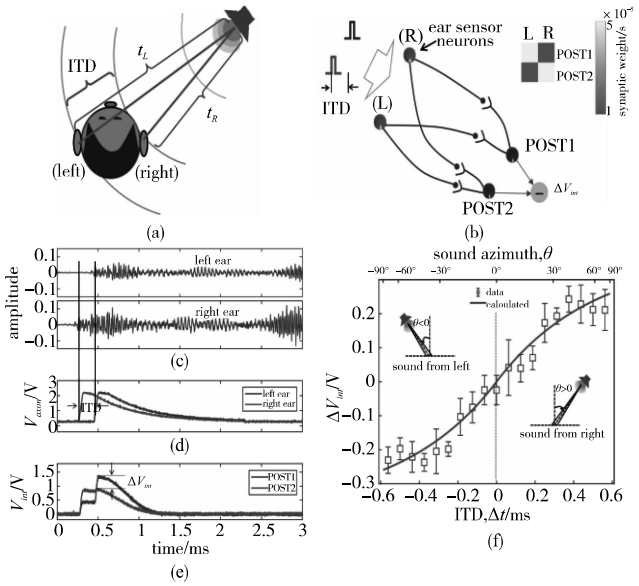


图26 脉冲神经网络测试

Fig.26 The test of SNN

4 结束语

尽管人们建立神经元模型[83]和发现 HEBB 学习规则[84](人脑中突触前神经元向连接突触后神经元的持续重复的刺激可以导致突触传递效能的增加或者减少) 已经超过 60 多年，但是至今仍未真正了解大脑实际运算机制。欣慰的是类脑计算的实现可以从模拟神经元形态网络的结构和功能着手，而无须等待完全弄清大脑的机制[9]。因此本文简要叙述神经形态器件的几个主要器件，并介绍脉冲神经网络电路和计算一存储一体的神经网络电路，作为目前神经形态器件与类脑神经网络电路基本现状综述。

需要指出的是，神经形态器件作为类脑神经网络最重要的部分，仍然有许多问题尚未解决。例如，人工突触通常采用忆阻器(也可称为连续可变电导的非易失性存储器 (NVM)) 来实现突触可塑性[85]，目前研究论文中常见的忆阻器件有: 相变存储器(PCM) [86]、阻变存储(RRAM) [87]以及多个并联式的二进制 NVM，如 CBRAM (导电桥接) [88]、丝状或非丝状 RRAM[89,90]等。而这些多值忆阻器件性能似乎还不太稳定[91]，其集成工艺也不成熟。特别是类脑神经网络有条件集成千亿个神经元器件，低压低功耗的性能要求成为基于多值忆阻器器件研究开发神经元器件的最重要指标[92]。因此，研究者开始探索纳米尺度的低功耗忆阻器，除了本文提到的石墨烯量子点忆阻器之外，如基于电子转移分子存储器[93]、基于电荷存储转移的石墨烯晶体管[94,95]等，这类器件不仅具有良好的低功耗特性，而且能够实现生物神经元所具备的存储与计算融合于一体的数据处理过程。另一方面，尽管当前多输入浮栅管能够实现权值求和，但阈值触发与不应期等功能仍采用复杂的 COMS 电路实现，因此纳米尺寸的神经元器件仍需要大量研究。

类脑神经网络可以说是人们研究和仿制大脑运算机制的初步成果，其中以模拟神经元结构为主的脉冲神经网络作为最重要的发展方向。近年来也出现了一些基于脉冲神经网络的神经形态处理器[96]，如 IBM 的 TrueNorth[97]、英特尔的 Loihi[98]和高通的 Zeroth 等芯片，但遗憾的是精确度还是无法与卷积神经网络(CNN) 及其递归神经网络(RNN) 混合组成的深度学习人工智能芯片相比[99]。而另一个研究方向是以模拟神经网络功能为主，为了更好地实现深度学习的并行计算，人们也设计出专用人工智能芯片，如 Intel 的 Nervana、谷歌的 TPU 以及国内中科院推出的寒武纪[100]等芯片，尽管引入多级缓存、多核/众核、并行可重构等并行计算技术，但都是以存储与计算分离的冯·诺依曼架构为基础，其存储和处理数据之间的传输存在瓶颈。

因此，要突破现行计算机架构的“存储墙”限制，类脑神经网络需要模拟生物网络、采用神经形态器件进行计算处理。与传统的人工神经网络不同，类脑神经网络在处理信息时，采用的是时间编码(temporal coding) ，这种编码能够携带时空、频率等信息，可以模拟各种神经信号，非常适合大脑神经元信号的处理，是进行处理计算复杂时空信号的有效工具[101]。并且通过模拟生物学习规则(STDP) ，最终实现非监督学习，更重要的是类脑计算以异步的、事件驱动的方式进行工作，更易于在硬件上实现计算与信息存储于一体的分布式运算，以此实现最接近人脑结构和处理信息的工作方式。有理由相信随着类脑计算的发展，将会引领人工智能走向新高度。

参考文献:

[1] 陶建华，陈云霁. 类脑计算芯片与类脑智能机器人发展现状与思考[J]. 中国科学院院刊，2016，31(7) : 803-811. (Tao Jianhua，Chen Yunji. Current status and consideration on brain-like computing chip and brain-like intelligent robot [J]. Bulletin of the Chinese Academy of Sciences，2016，31(7) : 803-811.)

[2] LeCun Y，Bengio Y，Hinton G. Deep learning [J]. Nature，2015，521: 436-444.

[3] Tenenbaum J B，Kemp C，Griffiths T L，*et al*. How to grow a mind: statistics structure and abstraction [J]. Science，2011，331(6022) : 1279-1285.

[4] 殷明惠，杨玉超，黄如. 神经形态器件现状与未来[J]. 国防科技，2016，37(6) : 23-30. (Yin Minghui，Yang Yuchao，Huang Ru. Progress and outlook of neuromorphic devices [J]. National Defense Science & Technology，2016，37(6) : 23-30.)

[5] Abbott L F，Nelson S B. Synaptic plasticity: taming the beast [J]. Nature Neuroscience，2000，3(11) : 1178-1183.

[6] Bi G Q, Poo M M. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type [J]. Journal of Neuroscience, 1998, 18(24): 10464–10472.
[7] 张晨曦,陈艳,仪明东,等. 基于忆阻器模拟的突触可塑性的进展研究[J]. 中国科学: 信息科学, 2018, 48(2): 115–142. (Zhang Chenxi, Chen Yan, Yi Mingdong, *et al*. Recent progress in memristors for stimulating synaptic plasticity [J]. Scientia Sinica Informationis, 2018, 48(2): 115–142.)
[8] He Wei, Huang Kejie, Ning Ning, *et al*. Enabling an integrated rate-temporal learning scheme on memristor [J]. Scientific Reports, 2015, 4: article No. 4755.
[9] Gopalakrishnan R, Basu A. Robust doublet STDP in a floating-gate synapse [C]//Proc of International Joint Conference on Neural Networks. Piscataway, NJ: IEEE Press, 2014: 4296–4301.
[10] Snider G S. Spike-timing-dependent learning in memristive nanodevices [C]//Proc of IEEE International Symposium on Nanoscale Architectures. Piscataway, NJ: IEEE Press, 2008: 85–92.
[11] Zhu Liqiang, Wan Changjin, Guo Liqiang, *et al*. Artificial synapse network on inorganic proton conductor for neuromorphic systems [J]. Nature Communications, 2014, 5: article No. 3158.
[12] Rajendran B, Liu Yong, Seo J S, *et al*. Specifications of nanoscale devices and circuits for neuromorphic computational systems [J]. IEEE Trans on Electron Devices, 2013, 60(1): 246–253.
[13] Wikipedia. Neuron [EB/OL]. https://en.wikipedia.org/wiki/Neuron.
[14] Gerstner W, Kistler W M. Spiking neuron models: single neurons, populations, plasticity [M]. Cambridge: Cambridge University Press, 2002.
[15] Mosleht S, Sahint C, Liut L, *et al*. An energy efficient decoding scheme for nonlinear MIMO-OFDM network using reservoir computing [C]//Proc of International Joint Conference on Neural Network. Piscataway, NJ: IEEE Press, 2016: 1166–1173.
[16] Danesh W, Zhao C, Wysocki B T, *et al*. Channel estimation in wireless OFDM systems using reservoir computing [C]//Proc of IEEE Symposium on Computational Intelligence for Security and Defense Applications. Piscataway, NJ: IEEE Press, 2015: 1–5.
[17] Reich D S, Mechler F, Purpura K P, *et al*. Interspike intervals, receptive fields, and information encoding in primary visual cortex [J]. Journal of Neuroscience, 2000, 20(5): 1964–1974.
[18] Feng J, Brown D. Integrate-and-fire models with nonlinear leakage [J]. Bulletin of Mathematical Biology, 2000, 62(3): 467–481.
[19] Gerstner W, Kistler M. Spiking neuron models: single neurons, populations, plasticity [M]. New York: Cambridge University Press, 2002: 110–133.
[20] Maass W, Bishop C M E. Pulsed neural networks [M]. Cambridge, MA: MIT Press, 1999.
[21] Nastschläger T, Ruf B. Spatial and temporal pattern analysis via spiking neurons [J]. Network: Computation in Neural Systems, 1998, 9(3): 319–332.
[22] Rao R P, Sejnowski T J. Spike-timing-dependent Hebbian plasticity as temporal difference learning [J]. Neural Computation, 2001, 13(10): 2221–2237.
[23] Gerstner W, Ritz R, Hemmen J L V, *et al*. Why spikes? Hebbian learning and retrieval of time-resolved excitation patterns [J]. Biological Cybernetics, 1993, 69(5): 503–515.
[24] Masquelier T, Guyonneau R, Thorpe S J, *et al*. Spike timing dependent plasticity finds the start of repeating patterns in continuous spike trains [J]. PLoS One, 2008, 3(1): e1377.
[25] Masquelier T, Guyonneau R, Thorpe S J, *et al*. Competitive STDP-based spike pattern learning [J]. Neural Computation, 2009, 21(5): 1259–1276.
[26] Young J M, Waleszczyk W J, Wang C, *et al*. Cortical reorganization consistent with spike timing but not correlation dependent plasticity [J]. Nature Neuroscience, 2007, 10(7): 887–895.
[27] Azghadi M R, Al-Sarawi S, Abbott D, *et al*. A neuromorphic VLSI design for spike timing and rate based synaptic plasticity [J]. Neural Networks, 2013, 45(9): 70–82.
[28] Pfister J P, Gerstner W. Triplets of spikes in a model of spike timing-dependent plasticity [J]. Journal of Neurophyscience, 2006, 26(38): 9673–9682.
[29] Wikipedia. Hopfield network [EB/OL]. https://en.wikipedia.org/wiki/Hopfield network.
[30] Wikipedia. Feedforward neural network [EB/OL]. https://en.wikipedia.org/wiki/Feedforward neural network.
[31] Wikipedia. Self-organizingmap [EB/OL]. https://en.wikipedia.org/wiki/Self-organi-zing map.
[32] Indiveri G, Chicca E, Douglas R. A VLSI array of low-power spiking neurons and bistable synapses with spike-timing dependent plasticity [J]. IEEE Trans on Neural Networks, 2006, 17(1): 211–221.
[33] Indiveri G, Corradi F, Qiao N. Neuromorphic architectures for spiking deep neural networks [C]//Proc of IEEE International Electron Devices Meeting. Piscataway, NJ: IEEE Press, 2015: 68–71.
[34] Eryilmaz S B, Kuzum D, Yu S, *et al*. Device and system level design considerations for analog-nonvolatile-memory based neuromorphic architectures [C]//Proc of IEEE International Electron Devices Meeting. Piscataway, NJ: IEEE Press, 2015.
[35] Yu Shimeng, Kuzum D, Wong H S P. Design considerations of synap-tic device for neuromorphic computing [C]//Proc of IEEE International Symposium on Circuits and Systems. Piscataway, NJ: IEEE Press, 2014: 1062–1065.
[36] Jackson B L, Rajendran B, Corrado G S, *et al*. Nanoscale electronic synapses using phase change devices [J]. ACM Journal on Emerging Technologies in Computing Systems, 2013, 9(2): article No. 12.
[37] Chua L O. Memristor-the missing circuit element [J]. IEEE Trans on Circuit Theory, 1971, 18(5): 507–519.
[38] Shibata T, Ohmi T. An intelligent MOS transistor featuring gate-level weighted sum and threshold operation electron devices meeting [C]// Proc of International Electron Devices Meeting. Piscataway, NJ: IEEE Press, 1992: 919–922.
[39] Kohno A, Murakami H, Ikeda M, *et al*. Memory operation of silicon quantum-dot floating-gate metal-oxide-semicond-uctor field-effect transistors [J]. Japanese Journal of Applied Physics, 2001, 40(7): 721–723.
[40] Rahimi K, Diorio C, Hernandez C, *et al*. A simulation model for floating-gate MOS synapse transistors [C]//Proc of IEEE International Symposium on Circuits and Systems. Piscataway, NJ: IEEE Press, 2002.
[41] Smith A W, McDaid L J. A compact spike timing dependent plasticity circuit for floating gate weight implementation [J]. Neurocomputing, 2014, 124: 210–217.
[42] Shibata T, Ohmi T. A functional MOS transistor featuring gate-level weighted sum and threshold operations [J]. IEEE Trans on Electron Devices, 1992, 39(6): 1444–1455.
[43] Guo X, Bayat F M, Prezioso M, *et al*. Temperature-insensitive analog vector-by-matrix multiplier based on 55 nm nor flash memory cells [C]//Proc of IEEE Custom Integrated Circuits Conference. Piscataway, NJ: IEEE Press, 2017: 1–4.
[44] Mead C. Analog VLSI and neural systems [M]//[S. l.]: Addison-Wesley Longman Publishing, 1989.
[45] Sarpeshkar R. Analog versus digital: Extrapolating from electronics to neurobiology [J]. Neural Computation, 1998, 10(7): 1601–1638.
[46] Bayat F M, Guo X J, Klachko M. High-performance mixed-signal neurocomputing with nanoscale floating-gate memory cell arrays [J]. IEEE Trans on Neural Networks and Learning Systems, 2018, 29(10): 4782–4790.
[47] Hasler J, Marr B. Finding a roadmap to achieve large neuromorphic hardware systems [J]. Frontiers in Neuroscience, 2013, 7: 118.
[48] Bayat F M, Guo X, Om'mani H A, *et al*. Redesigning commercial floating-gate memory for analog computing applications [C]//Proc of IEEE International Symposium on Circuits and Systems. Piscataway, NJ: IEEE Press, 2015: 1921–1924.
[49] Schlottmann C R, Hasler P. A highly dense, low power, programmable analog vector-matrix multiplier: the FPAA implementation [J]. IEEE Journal on Emerging and Selected Topics Circuits Systems, 2011, 1(3): 403–411.
[50] Strukov D B, Snider G S, Stewart D R, *et al*. The missing memristor found [J]. Nature, 2008, 453(7191): 80–83.
[51] 杨玖,王丽丹,段书凯. 一种反向串联忆阻突触电路的设计及应用[J]. 中国科学: 信息科学, 2016, 46(3): 391–403. (Yang Jiu, Wang Lidan, Duan Shukai. An anti-series memristive synapse circuit design and its application [J]. Scientia Sinica Informationis, 2016, 46(3): 391–403.)
[52] Strukov D B, Williams R S. Exponential ionic drift: fast switching and low volatility of thin-film memristors [J]. Applied Physics A, 2009, 94(3): 515–519.
[53] Pershin Y V, Ventra M D. Memory effects in complex materials and nanoscale systems [J]. Advances in Physics, 2011, 60(2): 145–227.
[54] Wong H S P, Lee H Y. Metal-oxide RRAM [J]. Proceeding of the IEEE, 2012, 100(6): 1951–1970.
[55] Lee H Y, Chen P S, Wu T, *et al*. HfOx bipolar resistive memory with robust endurance using AlCu as buffer electrode [J]. IEEE Elec–

tron Device Letters, 2009, 30(7): 703–705.
[56] Lee H Y, Chen P S, Wu T Y, *et al*. Low power and high speed bipolar switching with a thin reactive Ti buffer layer in robust HfO2 based RRAM[C]//Proc of IEEE International Electron Devices Meeting. Piscataway, NJ: IEEE Press, 2008: 1–4.
[57] Govoreanu B, Kar G S, Chen Y Y, *et al*. 10 × 10 nm^2 Hf/HfOx crossbar resistive RAM with excellent performance, reliability and low-energy operation [C]//Proc of International Electron Device Meeting. Piscataway, NJ: IEEE Press, 2011: 729–732.
[58] Eryilmaz S B, Kuzum D, Jeyasingh R, *et al*. Brain-like associative learning using a nanoscale nonvolatile phase change synaptic device array[J]. Frontiers Neuroscience, 2014, 8: 205.
[59] Yu Shimeng, Gao Bin, Fang Zheng, *et al*. A low energy oxide-based electronic synaptic device for neuromorphic visual systems with tolerance to device variation[J]. Advanced Materials, 2013, 25(12): 1774–1779.
[60] Mnih V, Kavukcuoglu K, Sliver D, *et al*. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529–533.
[61] Li Lingling, Wu Gehui, Yang Guohai, *et al*. Focusing on luminescent graphene quantum dots: current status and future perspectives [J]. Nanoscale, 2013, 5(10): 4015–4039.
[62] Wang Changhong, He Wei, Tong Yishu, *et al*. Memristive devices with highly repeatable analog states boosted by graphene quantum dots [J/OL]. Small, 2017, 13(20). https://doi.org/10.1002/smll.201603435.
[63] Morie T, Matsuura T, Nagata M, *et al*. A multinanodot floating-gate MOSFET circuit for spiking neuron models [J]. IEEE Trans on Nanotechnology, 2003, 2(3): 158–164.
[64] Liu S C, Rico M. Temporally learning floating-gate VLSI synapses [C]//Proc of IEEE International Symposium on Circuits and Systems. Piscataway, NJ: IEEE Press, 2008: 2154–2157.
[65] Diorio C, Hasler P, Minch B A, *et al*. A single-transist-or silicon synapse [J]. IEEE Trans on Electron Devices, 1996, 43(11): 1972–1980.
[66] Hasler P. Foundations of learning in analog VLSI[D]. Pasadena: California Institute of Technology, 1997.
[67] Arthur J, Boahen K. Learning in silicon: timing is everything[C]// Advances in Neural Information Processing Systems. Piscataway, NJ: IEEE Press, 2006: 75–82.
[68] Song S, Miller K, Abbott L. Competitive Hebbian learning through spike-timing-dependent synaptic plasticity [J]. Nature Neuroscience, 2000, 3(9): 919–926.
[69] Wang H X, Gerkin R C, Nauen D W, *et al*. Coactivation and timing-dependent integration of synaptic potentiation and depression[J]. Nature Neuroscience, 2005, 8(2): 187–193.
[70] Froemke R C, Tsay I A, Raad M, *et al*. Contribution of individual spikes in burst-induced long-term synaptic modi-fication[J]. Journal of Neurophysiology, 2006, 95(3): 1620–1629.
[71] Sjöström P J, Turrigiano G G, Nelson S B. Rate, timing, and cooperativity jointly determine cortical synaptic plasticity [J]. Neuron, 2001, 32(6): 1149–1164.
[72] Gopalakrishnan R, Basu A. Triplet spike time-dependent plasticity in a floating-gate synapse [J]. IEEE Trans on Neural Networks, 2017, 28(4): 778–790.
[73] Shukla A, Prasad S, Lashkare S, *et al*. A case for multiple and parallel RRAMs as synaptic model for training SNNs[C]//Proc of International Joint Conference on Neural Networks. Piscataway, NJ: IEEE Press, 2018: 1–8.
[74] Rossum M, Bi G Q, Turrigiano G G, *et al*. Stable Hebbian learning from spike timing-dependent plasticity [J]. The Journal of Neuroscience, 2000, 20(23): 8812–8821.
[75] Zamarrenoramos C, Camunasmesa L A, Perez-Carrasco J A, *et al*. On spike-timing-dependent-plasticity, memristive devices, and building a self-learning visual cortex [J]. Frontiers Neuroscience, 2011, 5: 1–22.
[76] Wang W, Pedretti G, Milo V, *et al*. Learning of spatiotemporal patterns in a spiking neural network with resistive switching synapses [J]. Science Advances, 2018, 4(9): 4752.
[77] Pedretti G, Milo V, Ambrogio V, *et al*. Memristive neural network for on-line learning and tracking with brain-inspired spike timing dependent plasticity[J]. Scientific Reports, 2017, 7(1): article No. 5288.
[78] Ambrogio S, Ciocchini N, Laudato M, *et al*. Unsupervised learning by spike timing dependent plasticity in phase change memory(PCM) synapses[J]. Frontiers Neuroscience, 2016, 10: 56.
[79] Sebastian A, Tuma T, Papandreou T, *et al*. Temporal correlation detection using computational phase-change memory[J]. Nature Communication, 2017, 8(1): article No. 1115.
[80] Wang Zhongqiang, Ambrogio S, Balatti S, *et al*. A 2-transistor/1-resistor artificial synapse capable of communication and stochastic learning in neuromorphic systems[J]. Frontiers Neuroscience, 2015, 8: 438.
[81] Xie Xiurui, Qu Hong, Liu Guisong, *et al*. Efficient training of supervised spiking neural networks via the normalized perceptron based learning rule[J]. Neurocomputing, 2017, 241: 152–163.
[82] Hancock K E, Delgutte B. A physiologically based model of interaural time difference discrimination[J]. Journal of Neuroscience, 2004, 24(32): 7110–7117.
[83] McCulloch W S. Pitts W. A logical calculus of the ideas immanent in nervous activity [J]. Bulletin of Mathematical Biology, 1990, 52(4): 99–115.
[84] Hebb D O. The organization of behavior: a neuropsychological theory [M]. [S.l.]: Psychology Press, 2005.
[85] Geoffrey W B, Shelby R M, Sebastian A, *et al*. Neuromorphic computing using non-volatile memory [J]. Advances in Physics: X, 2017, 2(1): 89–124.
[86] Jackson B L, Rajendran B, Corrado G S, *et al*. Nanoscale electronic synapses using phase change devices [J]. ACM Journal on Emerging Technologies in Computing Systems, 2013, 9(2): 1–20.
[87] Waser R, Regina D, Staikov G, *et al*. Redox-based resistive switching memories nanoionic mechanisms, prospects, and challenges[J]. Advanced Materials, 2009, 21: 2632–2663.
[88] Kim K H, Gaba S, Wheeler D, *et al*. A functional hybrid memristor crossbar-array/CMOS system for data storage and neuromorphic applications[J]. Nano Letters, 2012, 12: 389–395.
[89] Valov I, Waser R, Jameson J R, *et al*. Electrochemical metallization memories-fundamentals, applications, prospects [J]. Nanotechnology, 2011, 22(25): 254003.
[90] Seo K, Kim I, Jung S, *et al*. Analog memory and spiketiming-dependent plasticity characteristics of a nanoscale tita-nium oxide bilayer resistive switching device[J]. Nanotechnology, 2011, 22(25): 254023.
[91] Tosson A M S, Yu S, Anis M H, *et al*. A study of the effect of RRAM reliability soft errors on the performance of RRAM-based neuromorphic systems[J]. IEEE Trans on Very Large Scale Integration Systems, 2017, 25(11): 3125–3137.
[92] Benjamin B V, Gao P, McQuinn E, *et al*. Neurogrid: a mixed-analog digital multichip system for large-scale neural simulations [J]. Proceedings of the IEEE, 2014, 102(5): 699–716.
[93] Lindsay S M, Ratner M A. Molecular transport junctions: clearing mists[J]. Advanced Materials, 2007, 19(1): 23–31.
[94] Hu J X, Stecklein G, Anugrah Y, *et al*. Using programmable graphene channels as weights in spin-diffusive neuromorphic computing [J]. IEEE Journal on Exploratory Solid-State Computational Devices and Circuits, 2018, 4(1): 26–34.
[95] Mohammad T S, Du Y, Torres J, *et al*. Low-power electrochemically tunable graphene synapses for neuromorphic computing [J]. Advanced Materials, 2018, 30(36): 1802353.
[96] Nawrocki R A, Voyle R M, Shaheen S E. A mini review of neuromorphic architectures and implementations [J]. IEEE Trans on Electron Devices, 2016, 6(10): 3819–3829.
[97] Akopyan F, Sawada J, Cassidy A, *et al*. TrueNorth: design and tool flow of a 65 mW 1 million neuron programmable neurosynaptic chip [J]. IEEE Trans on Computer-Aided Design of Integrated Circuits and Systems, 2015, 34(10): 1537–1557.
[98] Davies M, Srinivasa N, Lin T H, *et al*. Loihi: a neuromorphic manycore processor with on-chip learning[J]. IEEE Micro, 2018, 38(1): 82–99.
[99] Saeed M, Khan A A, Kamboh A M. Comparison of classifier architectures for online neural spike sorting[J]. IEEE Trans on Neural Systems and Rehabilitation Engineering, 2017, 25(4): 334–344.
[100] Yin Shouyi, Ouyang Peng, Tang Shibin, *et al*. A high energy efficient reconfigurable hybrid neural network processor for deep learning applications[J]. IEEE Journal of Solid-State Circuits, 2018, 53(4): 968–982.
[101] 蔺想红, 王向文, 张宁, 等. 脉冲神经网络的监督学习算法研究综述[J]. 电子学报, 2015, 43(3): 576–586. (Lin Xianghong, Wang Xiangwen, Zhang Ning, *et al*. Supervised learning algorithms for spiking neural networks [J]. Acta Electronica Sinica, 2015, 43(3): 576–586.)