文章编号:1007 - 4252(2024)02 - 0002 - 104　　DOI: 10.20027/j.gncq.2024.0015

基于 Ag/CeO_2 /ITO 忆阻器的片上学习神经网络

胡俊达[1],刘　昊[1,2],徐　辛[1,2],黄佳俊[3],杨　峰[1],张　勇[1,*]

(1. 西南交通大学电气工程学院,磁浮技术与磁浮列车教育部重点实验室,成都 610031;2. 西南交通大学物理科学与技术学院,成都 610031;3. 西南交通大学材料科学与工程学院,成都 610031)

摘要:现有的忆阻神经网络存在学习速率慢、精度较低、电路复杂等问题,为了实现标准、高效的片上学习,设计了基于 Ag/CeO_2/ITO 忆阻器的片上学习神经网络。由一个晶体管和两个 Ag/CeO_2/ITO 忆阻器作为突触神经元,该突触结构具有更大的权值范围,Ag/CeO_2/ITO 忆阻器的阈值特性简化了非破坏性读取的步骤和电路结构,神经网络通过四种电路单元实现无数模转换和无外部处理器的片上学习,每层神经网络可以在两个时钟周期内并行更新权值,该方法避免了由数据传输所造成的延迟、功耗和误差。最后,仿真验证设计的忆阻神经网络并应用于字符图像和鸢尾花识别,识别准确率均能达到 95% 以上,且忆阻器件的差异性对准确率的影响不大,证明其具有有效性和鲁棒性。

关键词:忆阻器;片上学习;神经网络;阈值;识别

中图分类号:TP343、TP183　　**文献标志码**:A

On-chip Learning Neural Networks with Ag/CeO_2 /ITO Memristors

HU Junda[1], LIU Hao[1,2], XU Xin[1,2], HUANG Jiajun[3], YANG Feng[1], ZHANG Yong[1]

(1. School of Electrical Engineering, Key Laboratory of Magnetic Levitation Technologies and Maglev Trains, Ministry of Education, Southwest Jiaotong University, Chengdu 610031, China.

2. School of Physical Science and Technology, Southwest Jiaotong University, Chengdu, 610031, China.

3. School of Materials Science and Engineering, Southwest Jiaotong University, Chengdu, 610031, China)

Abstract: The existing memristor neural network has the problems of slow learning rate, low accuracy and complex circuit. In order to achieve standard and efficient on-chip learning, an on-chip learning neural network based on Ag/CeO_2/ITO memristor is designed. A transistor and two Ag/CeO_2/ITO memristors are used as synaptic neurons, and the synaptic structure has a larger weight range. The threshold characteristics of Ag/CeO_2/ITO memristor simplify the steps and circuit structure of non-

收稿日期:2024-03-07;　　**修订日期**:2024-03-28

基金项目:重点研发计划国家磁约束核聚变专项基金(2022YFE03020002)

作者简介:胡俊达(1998-),男,硕士,主要研究方向为忆阻器应用(E-mail:: hjd2021200431@my.swjtu.edu.cn).

通信作者:张　勇(1969-),女,博士,教授,主要研究方向为超导材料、半导体材料(E-mail:yongzhang@swjtu.edu.cn).

destructive reading. Each layer of the neural network can update the weights in parallel in two clock period. This method avoids the delay, power consumption and error caused by data transmission. Finally, the proposed memristor neural network is simulated and applied to the recognition of character images and iris flowers. The recognition accuracy can reach more than 95%, and the difference of memristor devices has little impact on the accuracy, which proves the effectiveness and robustness of the proposed memristor neural network.

Key words: Memristor; On-chip learning; Neural network; Threshold; recognition

0　引言

神经网络因其强大的学习能力和鲁棒性而在分类、回归、聚类、优化等任务中被广泛应用[1]。然而,随着神经网络层数的增加,计算量也会迅速增加,传统的冯诺依曼架构难以满足神经网络巨大的运算需求。忆阻器[2]是一种具有记忆功能的非易失性存算一体器件,并且具有高度集成化、与 CMOS 兼容等优势,使得其可以模拟神经突触,以交叉棒阵列的形式构成神经网络,这种忆阻交叉阵列可以快速完成乘加运算,从而应对大规模的计算量[3-5]。忆阻阵列的学习方式可以分为两类,片外学习和片上学习。片外学习是指在外部计算机上对忆阻神经网络进行训练,然后将训练好的权重写入其中。这种方法结构简单、可以使用任何激活函数,但是容易受到器件差异性的影响,而且频繁的数据传输和数模转换会对神经网络的性能(精度、延时和能耗等)产生很大影响。片上学习是指在忆阻交叉阵列上直接执行训练算法,利用反馈信号来调整权重,从而减小器件变异和数据传输带来的影响,但片上学习较为复杂和困难,如何实现忆阻神经网络的片上学习是一个亟需解决的问题。

近年来,学者们对片上学习的简化算法和电路设计进行了大量的研究。Adhikari 等人[6]利用随机权重变化算法(RWC)来更新忆阻神经网络的权重,虽然该算法会增加迭代的次数,但是由于没有复杂的反向传播以及导数计算,每次迭代所需的时间得以缩短。王春华等人[7]提出了一种基于权重同时扰动算法(WSP)的神经网络电路架构,在训练周期内增加了一个操作步骤,但是考虑了误差变化的方向和幅度,它比 RWC 有更快的收敛速率。徐聪等人[8]改进了 WSP 算法,将加权同时扰动算法(WSSP)应用于忆阻神经网络,该算法只随机改变权重和。李博勋等人[9]设计了一个混合训练框架,使用两个忆阻交叉阵列分别用于前向传播和反向传播,该框架可以实现误差反向传播,但也增加了电路尺寸,而且由于忆阻器件的差异,第二个忆阻交叉阵列不能准确地复制第一个忆阻交叉阵列的权重值。上面这些学习方法通过简化算法或电路结构使之更易实现,但也会牺牲计算精度和收敛速度。为了保证神经网络的性能,Soudry 等人[10]设计了一个集成正向传播和反向传播的忆阻交叉阵列,它的突触由两个晶体管和一个忆阻器(2T1M)组成,并在一个周期内控制两个晶体管的通断以实现标准的监督学习,但其 2T1M 的突触结构较为复杂且突触权值只能为正,作者也未对外围电路进行设计。Hasan 等人[11]采用两个忆阻器作为突触(2M),该结构具有更高的精度和更小的尺寸,并且仿真结果表明其功耗更低,但是其权重调整的操作较为复杂耗时,只能逐列调整突触权重,还需要数模互换和外部存储器。

因此现有的忆阻神经网络存在学习速率慢、精度较低、电路复杂等问题,为了解决上述问题,实现高效、标准的片上学习,本研究设计了基于 Ag/CeO_2/ITO 忆阻器的片上学习神经网络。采用 1T2M 突触神经元,并利用 Ag/CeO_2/ITO 忆阻器的阈值特性简化了非破坏性读取步骤和电路结构。设计的忆阻神经网络通过四个电路单元实现梯度下降算法,训练过程在两个时钟周期内并行完成,并且不需要外部处理器和数模互换,仿真表明,该忆阻神经网络能很好地完成各类识别任务,具有有效性和鲁棒性。

1　Ag/CeO_2/ITO 忆阻器

忆阻器最早由 Leon Chua 于 1971 年在理论上提出[12],2008 年惠普实验室制备了 TiO_2 忆阻器的物理器件[13]。根据氧空位迁移模型,当施加的脉冲电压持续时间足够短时,可以近似认为忆阻器阻值的变化与电压幅值 Vplus 和持续时间 Tplus 的乘积成正比:

$$\Delta M = \frac{\mu_V M_{min}(M_{min} - M_{max}) f(s(t)) V_{plus} T_{plus}}{M D^2} \quad (1)$$

式中，M_{max}和 M_{min}分别为忆阻器的最大和最小阻值，s(t)为掺杂层的厚度占比，μ_V为迁移率，f(s(t))为窗函数，M 为忆阻器当前阻值，D 为忆阻器的厚度。

除了 TiO_2，WO_X[14]、AlO_X[15]、HfO_X[16]、Ta_2O_5[17]、Cu_2O[18]等材料也可以制成模拟型忆阻器件，本文采用磁控溅射的方法制成 Ag/CeO_2/ITO 忆阻器，CeO_2溅射时间为 30 min，图 1(a)是 Ag/CeO_2/ITO 忆阻器结构示意图和扫描电镜图，CeO_2层厚度约 150 nm。利用电化学工作站所测得循环电压下的 I-V 曲线如图 1(b)，脉冲电压下的电流曲线如图 1(c)。

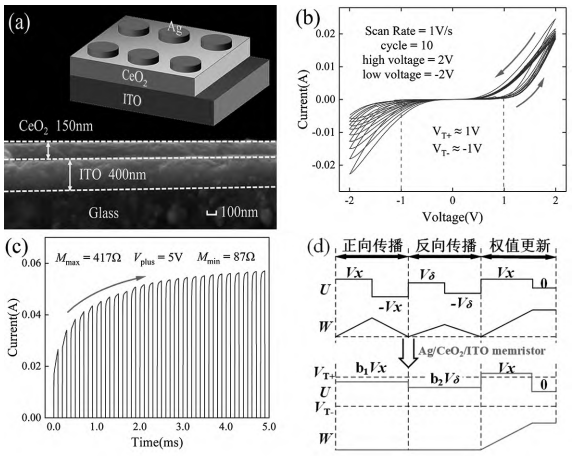

图 1 (a) Ag/CeO_2/ITO 忆阻器扫描电镜图和示意图；(b)循环电压下的 I-V 曲线；(c)脉冲电压下的电流曲线；(d)非破坏性读取步骤

Fig. 1 (a) SEM and schematic diagram of Ag/CeO_2/ITO memristor; (b) I-V curve at cyclic voltage; (c) Current curve at pulsed voltage; (d) Non-destructive reading steps

从图 1(b)可以看出该忆阻器具有一定阈值特性，当电压在-1V ~ 1V 之间时，其阻值基本不变，只有当电压超过±1V 时，阻值才会发生明显变化。图 1(c)显示在脉冲作用下，其电阻从 87Ω 到 417Ω 连续变化，证明了忆阻器具有良好的模拟性能。图 1(d)为不同忆阻器件非破坏性读取的步骤，对于无阈值特性的忆阻器，需要施加相同时间的反向电压来实现非破坏性读取，而对于此 Ag/CeO_2/ITO 忆阻器，只需采用不超出阈值的电压进行读取，高于阈值的电压用于调整阻值，此举可以简化操作步骤以及电路结构。

本文采用 Ag/CeO_2/ITO 忆阻器进行神经网络的搭建和仿真，利用氧空位迁移模型对其进行建模，具体参数见表 1。

表 1 Ag/CeO_2/ITO 忆阻器的模型参数

Table 1 Model parameters of the Ag/CeO_2/ITO memristor

参数	数值
最大阻值 M_{max}/Ω	417
最小阻值 M_{min}/Ω	87
正向阈值 V_{T+}/V	1
反向阈值 V_{T-}/V	-1
迁移率 μ_V/ $m^2 s^{-1} V^{-1}$	3e-12
厚度 D/m	1.5e-7
窗函数 f	Biolek
指数 p	2

2 突触神经元

突触神经元的结构具有多种形式，最简单的突触神经元由单个忆阻器(1M)构成[19]，该方案具有结构简单，电路尺寸小等优点，但会产生潜通路问题[20]，一种常用的解决方法是加入晶体管[21](1T1M)消除反向通路。然而这两种结构都只能实现正数的突触权重，在一定程度上限制了忆阻神经网络的性能，张洋[22]通过一个负的偏置值将突触权重拉低来实现正、负、零的权重，但本质上其权重范围并没有发生变化。为了兼顾潜通路、权重范围、电路尺寸等因素，本文采用如图 2(a)的 1T2M 突触结构，V_C是 NMOS 管的栅极控制电压。文中采用的 MOS 管均为 CMOS，因为 CMOS 管可以控制双极性信号[23]，神经元的输出满足：

$$V_Y = W V_X = R_f(G_+ - G_-) V_X = R_f\left(\frac{1}{M_+} - \frac{1}{M_-}\right) V_X \quad (2)$$

式中，V_X和 V_Y分别为神经元的输入和输出对应的电压，W 是突触权重值，G_+和 G_-分别为正向忆阻器和反向忆阻器的电导值，M_+和 M_-分别为正向忆阻器和反向忆阻器的电阻值。

与突触结构相连的运放共模输入电压为 0V，因此采用双电源供电，考虑失调电压时，神经元的输出为：

$$V_Y = R_f\left(\frac{1}{M_+} - \frac{1}{M_-}\right)V_X + \left(\frac{R_f}{M_+} + \frac{R_f}{R_0}\right)V_{OS1} - \left(\frac{R_f}{M_-} + \frac{R_f}{R_0} + 1\right)V_{OS2} \quad (3)$$

式中,V_{OS1}为与正向忆阻器相连的运放失调电压,V_{OS2}为与反向忆阻器相连的运放失调电压。从(3)中看到,运放的失调电压会对输出电压值产生影响,但由于失调电压一般为 mV 级或 μV 级,故对输出电压值的影响可以忽略。为了使 NMOS 管工作在线性区,需满足:

$$V_{GD} = V_C - V_X > V_{th} \quad (4)$$

式中,V_{GD}为栅漏电压,V_{th}为 NMOS 管的阈值电压。忆阻器 M_+和忆阻器 M_-均采用 Ag/CeO_2/ITO 忆阻器模型,且两忆阻器极性相反,当输入电压大于忆阻器阈值时,两忆阻器阻值改变且变化方向相反,图 2(b)展示了正负编程电压对突触权值的编程作用,除边界处存在非线性之外,权值变化基本与编程电压和时间成正比,突触权值范围为$R_f\left(\frac{1}{M_{max}} - \frac{1}{M_{min}}\right) \sim R_f\left(\frac{1}{M_{min}} - \frac{1}{M_{max}}\right)$,而单忆阻器突触结构的权值范围为$R_f\frac{1}{M_{max}} \sim R_f\frac{1}{M_{min}}$,因此精度比单忆阻器突触高一倍。

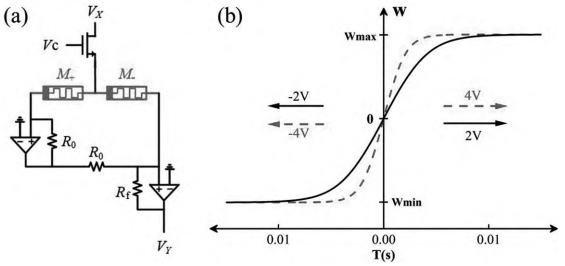

图 2　(a) 1T2M 突触神经元结构;(b) 正负编程电压对突触权值的更新过程

Fig. 2　(a) 1T2M Synaptic neuron structure; (b) The updating process of synaptic weights by positive and negative programming voltages

3　忆阻神经网络

3.1　控制信号

设计的忆阻神经网络主要有三个时序信号:读信号 Vr、写信号 Vw 和三角波 Vs,Vr、Vw 为脉冲信号,它们将一个训练周期分为四个阶段,如图 3,在 T1 阶段进行正向传播;在 T2 阶段进行误差反向传播;在 T3 和 T4 阶段内,进行权值更新,Vs 用于将误差量转换为 NMOS 管的导通时间。三种信号的周期满足:

$$T_r = T_s = 2\,T_w = T \quad (5)$$

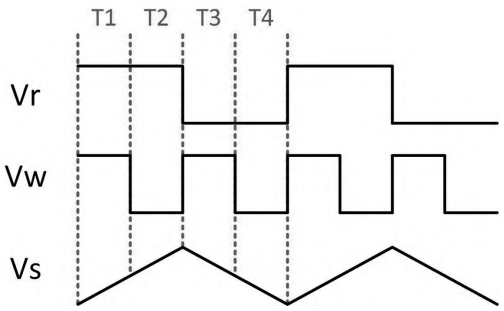

图 3　Vr、Vw、Vs 信号时序图

Fig. 3　Timing diagram of Vr, Vw and Vs signals

3.2　电路单元

图 4 展示了每层忆阻神经网络的结构,由忆阻交叉阵列和四种电路单元(前置单元、反向传播单元 BP、后置单元以及脉冲信号调制单元 PWM)构成,X、Y、Z、δ 和 e 分别是神经网络的输入值、输出值、激活值、输出误差和激活误差,它们对应的电压分别为 V_X、V_Y、V_Z、V_δ、V_e。i=1,2,…,m,m 为神经网络行数,j=1,2,…,n,n 为神经网络列数。

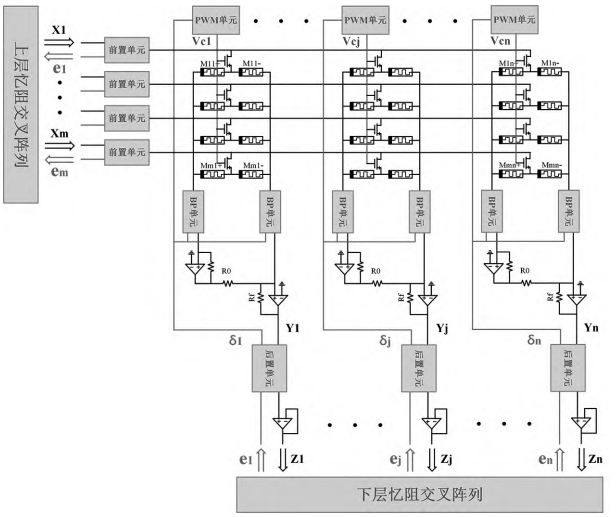

图 4　每层忆阻神经网络的结构

Fig. 4　The structure of the memristor neural network at each layer

前置单元位于神经网络的输入处,用于在 T1、T3、T4 阶段提供不同输入电压。BP 单元位于神经网络的输出前,用于在 T2 阶段内提供反向传播的误

差信号。后置单元位于神经网络的输出与激活函数之间,用于在 T2 ~ T4 阶段内保持 T1 阶段内的输出信号,此外该单元还承担计算激活函数导数的任务。PWM 单元通过接受列神经元的误差量,并配合 Vs 将其转变为占空比与之成正比的 PWM 信号,以控制 T3、T4 阶段内 NMOS 管的导通时间。激活函数使用了易于硬件实现的限幅函数,通过在输出端级联一个电压跟随器实现:

$$V_Z = f(V_Y) = \begin{cases} V_{SS} & V_Y < V_{SS} < 0 \\ V_Y & V_{ss} \leqslant V_Y \leqslant V_{cc} \\ V_{DD} & V_Y > V_{DD} > 0 \end{cases} \tag{6}$$

式中,V_{SS}和 V_{DD}是运放的电源电压。

电路单元依靠信号选择器对信号进行选择,信号选择器由一个反相器和两个传输门构成,如图 5(a)所示,当选择信号 Vp 为正时端口 V1 被选通,Vp 为负时端口 V2 被选通,为方便起见,后面内容中以图 5(b)代替图 5(a)。图 6 为各电路单元的结构示意图,下面将以多层神经网络为例描述每个电路单元的作用。

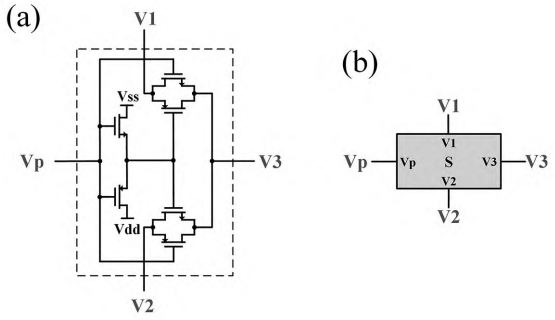

图 5 (a) 信号选择器;(b) 缩略图

Fig. 5 (a) Signal selector; (b) Thumbnail

在 T1 阶段,Vr 和 Vw 均为正,选择器 S1 ~ S7 均选通 V1 端口,对于前置单元,输入信号经过 OP1、OP2 比例缩小后输入神经网络,此举是为了在正向传播阶段实现非破坏性读取,缩小倍数为 R_0/R_1,满足:

$$V_{T-} \leqslant \frac{R_0}{R_1} V_{Xi} \leqslant V_{T+} \tag{7}$$

此时各层神经网络的激活值为:

$$Z^{(k)} = f^{(k)}(Y^{(k)}) = f^{(k)}\left(\frac{R_0}{R_1} W^{(k)} X^{(k)}\right) \tag{8}$$

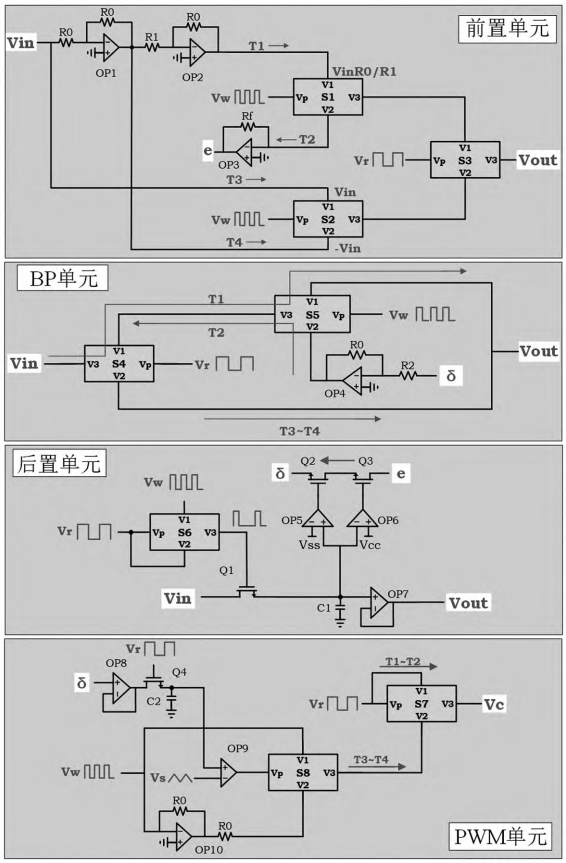

图 6 各电路单元的结构示意图

Fig. 6 Schematic diagram of the structure of each circuit unit

式中,上标(k)表示神经网络层数,k = 2, 3, …, L, L 是神经网络的总层数。为了保证信号的正向传播,其余单元需提供相应通路,信号通过 BP 单元的选择器 S4、S5 直接输出;后置单元中的选择器 S6 输出高电平,NMOS 管 Q1 导通,电容 C_1 充电至 $Y^{(k)}$;PWM 单元通过选择器 S7 输出高电平,以控制突触中的 NMOS 管保持导通。

在 T1 结束时,Vw 的下降沿关断后置单元中的 Q1,使得后置单元在接下来的 T2 ~ T4 阶段内保持正向传播得到的输出值 $Y^{(k)}$。

在 T2 阶段,Vr 为正,Vw 为负,选择器 S3、S4、S6、S7 选通 V1,选择器 S1、S2、S5 选通 V2,BP 单元的 OP4 将 T1 中产生的误差信号 δ 缩小后反向输入忆阻交叉阵列,缩小比例也需满足非破坏性读取的条件:

$$V_{T-} \leqslant \frac{R_0}{R_2} V_{\delta j \leqslant} > V_{T+} \tag{9}$$

突触中的忆阻器 M_+ 和 M_- 分别代表正负权值,

故两者需施加相反的误差电压:

$$V_{Mj+} = -V_{Mj-} = -\frac{R_2}{R_0}V_{\delta j} \tag{10}$$

误差信号在突触中产生的电流信号被前置单元中的运放 OP3 收集并转变为电压信号,此电压即为上一层神经网络激活值的误差量 $e^{(k-1)}$:

$$e^{(k-1)} = W^{(k)T}\delta^{(k)} \tag{11}$$

上层神经网络的后置单元会根据激活函数的导数判断是否允许 $e^{(k-1)}$ 输入:

$$\delta^{(k-1)} = e^{(k-1)} f^{(k-1)'}(Y^{(k-1)}) \tag{12}$$

当上层输出值 $Y^{(k-1)}$ 在 V_{SS} ~ V_{DD} 之间时,电压比较器 OP5、OP6 同时产生高电平,NMOS 管 Q3、Q4 同时导通,$e^{(k-1)}$ 才能作为上层的输出误差量 $\delta^{(k-1)}$ 进入 BP 单元和 PWM 单元,PWM 单元中的电容 C_2 通过 OP8 充电至 $\delta^{(k-1)}$,当 $Y^{(k-1)}$ 超出 V_{SS} 或 V_{DD} 时,即 $f^{(k-1)}(Y^{(k-1)}) = 0$,$e^{(k-1)}$ 被隔离,PWM 单元中的电容 C_2 通过 OP8 放电至 0。

后置单元的结构可以根据不同激活函数灵活改变,例如激活函数为 ReLU 时,只需使用一个电压比较运放和一个 NMOS 管即可实现输出大于零时导通,小于零时隔离。反向传播阶段内,PWM 单元也始终提供高电平。

T3 ~ T4 阶段为权值更新阶段,根据梯度下降算法,权值修正量为:

$$\Delta W = \eta\delta X^T \tag{13}$$

权重修正量与误差量和输入量的乘积成正比,故每个神经元的权值修正量均不相同,如果只使用脉冲电压编程,需要 m×n 次操作才能完成训练。为了加快电路的运行速度,采用脉冲电压和时间的混合编程方法,由于同一行的神经元具有相同的输入量,同一列的神经元具有相同的误差量,因此保持输入量不变,将每一列的误差量转化成 NMOS 管的导通时间,即可实现并行调整所有神经元的权重值。具体方案为:前置单元在 T3 阶段内提供原输入电压,在 T4 阶段内提供反向输入电压,当列误差量为正时,PWM 单元控制突触内的 NMOS 管在 T3 阶段内导通,当列误差量为负时,控制突触内的 NMOS 管在 T4 阶段内导通,导通时间 Δt 与输出误差量 $\delta^{(k-1)}$ 成正比。表 2 列出了编程过程中可能出现的四种情况。

表 2　编程中的四种情况

Table 2　Four cases in programming

X	前置单元输出电压		δ	MOS 管导通时间		ΔW
	T3	T4		T3	T4	
+	+	−	+	Δt	0	+
+	+	−	−	0	Δt	−
−	−	+	+	Δt	0	−
−	−	+	−	0	Δt	+

假定三角波 Vs 的幅值为 Vm,容易推得 Δt:

$$\Delta t = \begin{cases} \dfrac{T}{4V_m} stp(V_\delta)\ (T3) \\ \dfrac{T}{4V_m} stp(-V_\delta)\ (T4) \end{cases} \tag{14}$$

根据(1),可以得到一个训练周期内忆阻器阻值的变化量:

$$\Delta M_{ij+} = \frac{\mu_V M_{min}(M_{min} - M_{max})T}{4D^2 M_{ij+} V_m} V_{\delta j} V_{Xi}$$

$$\Delta M_{ij-} = -\frac{\mu_V M_{min}(M_{min} - M_{max})T}{4D^2 M_{ij-} V_m} V_{\delta j} V_{Xi} \tag{15}$$

通过改变参数 Vm 和 T 可以调整神经网络的学习速率。在 T3、T4 阶段中,BP 单元始终保持直通,以保证每个忆阻器共地。

其中,stp(V_δ) 是 V_δ 的阶跃函数,$V_\delta > 0$ 时 stp(V_δ) 为 1,$V_\delta \leqslant 0$ 时 stp(V_δ) 为 0。

表 3 汇总了各电路单元在不同时间段内的状态,对于多层忆阻神经网络,除第一层无需 BP 单元外,其他三种电路单元电路应用于每一层神经网络。

利用 PSPICE 对忆阻神经网络进行仿真,参数为:Vr = −5 ~ 5 V,Vw = −5 ~ 5 V Vs = −3 ~ 3 V,T = 1 ms,R_0 = 1 kΩ,R_1 = 3 kΩ,R_2 = 3 kΩ,第一层 R_{f1} = 5 kΩ,第二层 R_{f2} = 0.5 kΩ,图 7 展示了三个周期内主要电路单元的波形,在第一个周期中,X 为正,δ 为负,忆阻器 M+阻值增加,M−阻值减小,突触权值减小;在第二个周期中,X 小于 V_{T+},故突触权值不变;第三周期与第一周期类似。

表 3　电路单元在各时间段内的状态

Table 3　States of circuit units in each time period

单元	T1	T2	T3	T4

前置单元	$\frac{R_1}{R_0}V_{in}$	产生 e	V_{in}	$-V_{in}$
BP 单元	直通	反向传播 δ	直通	直通
后置单元	直通	计算导数	保持	保持
PWM 单元	高电平	高电平	高低电平	高低电平

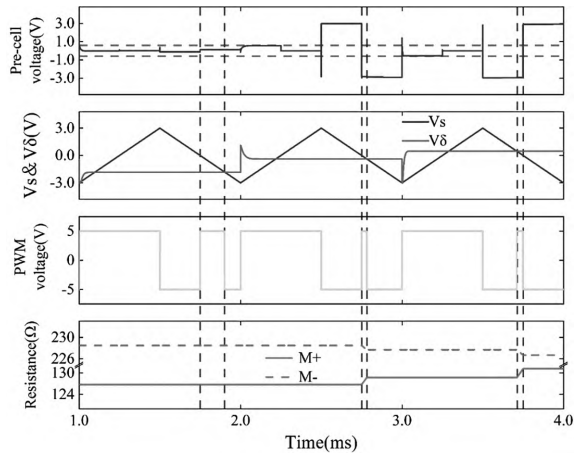

图 7 忆阻神经网络在 PSPICE 中仿真时的主要波形

Fig. 7 The dominant waveform of the memristor array when simulated in PSPICE

4 仿真与分析

在 MATLAB 和 PSPICE 中将设计的忆阻神经网络应用于光学字符识别(OCR)和鸢尾花识别(IRIS),以验证上述电路的有效性及鲁棒性。OCR 通过神经网络有效提取图像中的文字信息,为了方便起见,我们以 8 个 5 × 3 分辨率的字符("E, F, H, I, L, P, S, Z")为例,如图 8 所示,将字符的 15 个像素点二值化,黑白色像素点对应的输入电压分别为 3V、-3V,神经网络的结构为 15 × 12 × 8,将这八个训练样本以列向量的形式输入忆阻神经网络,训练完成后使用 8 个带有噪声的测试样本进行测试。IRIS 的神经网络结构为 4 × 12 × 3,将数据集打乱并归一化,输入电压范围为-3V ~3V,弗吉尼亚鸢尾花(setosa)、变色鸢尾花(versicolor)和维吉尼亚鸢尾花(virginica)对应的目标输出为(1,-1,-1),(-1, 1,-1),(-1,-1,1),选取 120 个样本进行训练,剩下的 30 的样本作为测试集。每个忆阻器的初始阻值都是随机的,由于忆阻器件之间存在差异性,其阻值变化率不同,为了测试忆阻神经网络的抗干扰能力,将忆阻器件分为理想器件和差异器件两组分别进行仿真,差异器件的阻值变化率满足正态分布。

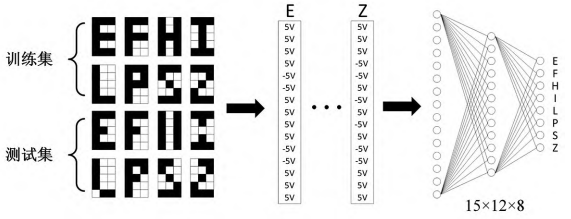

图 8 字符图像识别的流程

Fig. 8 The flow of character image recognition

图 9 展示了理想器件和差异器件在 OCR 和 IRIS 中的训练误差 MSE 图,可以看出无论器件是否存在差异,该忆阻神经网络都可以很好地学习对应任务,测试集的识别准确率如表 4,识别准确率为三次仿真的平均值,对于不同任务都有较高的识别准确率,且忆阻器件的差异性对识别准确率的影响不大,证明该忆阻神经网络具有有效性和鲁棒性。

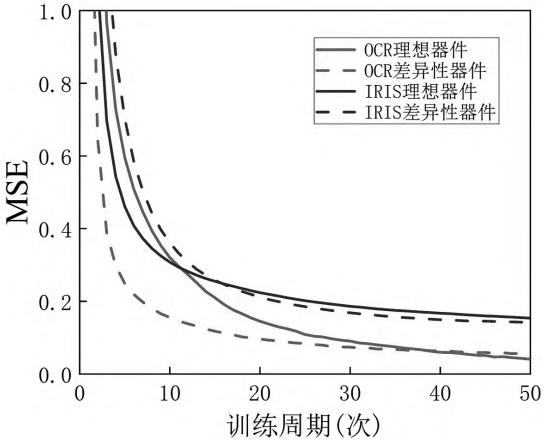

图 9 OCR 和 IRIS 的训练误差图

Fig. 9 Training error plots for OCR and IRIS

表 4 理想器件和差异器件的识别准确率

Table 4 Recognition accuracy of ideal and differential devices

识别准确率	OCR	IRIS
理想器件	100 %	96.76 %
差异器件	100%	95.56 %

对于忆阻神经网络,片上学习的功耗和延时主要取决于忆阻器的编程过程以及忆阻神经网络的规

模和结构。片外学习的功耗和延时主要取决于数据在忆阻神经网络和数据处理芯片之间的传输过程以及数据处理芯片的运算能耗和精度,相比于由忆阻神经网络复杂的结构所带来的延时和功耗,数据传输过程中产生的延时和功耗要大得多,因此,该忆阻神经网络能显著提高神经网络的学习效率。不过值得注意的是,由于 Ag/CeO_2/ITO 忆阻器自身的阻值较小,故会增加学习过程中的能耗,但是在读取阶段施加的小电压能大幅减小正向传播和反向传播中消耗的功率,而且读取操作可以在极短的时间内完成,因此能耗主要集中在权值更新阶段,而随着迭代的进行,权值的修正值会逐渐减小,因此功率也会逐渐减小。

5　结论

本研究利用 Ag/CeO_2/ITO 忆阻器搭建了片上学习的忆阻神经网络,Ag/CeO_2/ITO 忆阻器的阈值特性简化了操作和电路结构,1T2M 的突触结构具有更大的权值范围。忆阻阵列通过四种电路单元在不借助外部计算机和数模互换的情况下进行并行片上学习。对设计的忆阻神经网络进行仿真验证,仿真结果表明,该设计具有有效性和鲁棒性,可以很好地完成各类识别任务。该电路不仅适用于 Ag/CeO_2/ITO 忆阻器,还适用于其他带有阈值特性的模拟型忆阻器。

参考文献:

[1] Jain L C, Seera M, Lim C P, et al. A review of online learning in supervised neural networks [J]. Neural computing and applications, 2014, 25: 491-509.

[2] Qin X Z, Hu J D, Liu H, et al. Performance regulation of a ZnO/WOx - based memristor and its application in an emotion circuit [J]. The Journal of Physical Chemistry Letters, 2023, 14(12): 3039-3046.

[3] Hong Q H, Yan R, Wang C H, et al. Memristive circuit implementation of biological nonassociative learning mechanism and its applications [J]. IEEE Transactions on Biomedical Circuits and Systems, 2020, 14(5): 1036-1050.

[4] Yao P, Wu H Q, Gao B, et al. Fully hardware - implemented memristor convolutional neural network [J]. Nature, 2020, 577(7792): 641-646.

[5] Hu M, Chen Y, Yang J J, et al. A compact memristor-based dynamic synapse for spiking neural networks [J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2017, 36(8): 1353-1366.

[6] Adhikari S P, Kim H, Budhathoki R K, et al. A circuit-based learning architecture for multilayer neural networks with memristor bridge synapses [J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2015, 62(1): 215-223.

[7] Wang C H, Xiong L, Sun J, et al. Memristor-based neural networks with weight simultaneous perturbation training [J]. Nonlinear Dynamics, 2018, 95(4): 2893-2906.

[8] Xu C, Wang C H, Sun Y, et al. Memristor-based neural network circuit with weighted sum simultaneous perturbation training and its applications [J]. Neurocomputing, 2021, 462: 581-590.

[9] Li B, Wang Y Z, Wang Y, et al. Training itself: Mixed-signal training acceleration for memristor - based neural network [C]//2014 19th Asia and South Pacific Design Automation Conference (ASP - DAC). IEEE, 2014: 361-366.

[10] Soudry D, Di Castro D, Gal A, et al. Memristor-based multilayer neural networks with online gradient descent training [J]. IEEE Transactions on Neural Networks and Learning Systems, 2015, 26(10): 2408-21.

[11] Hasan R, Taha T M, Yakopcic C. On-chip training of memristor crossbar based multi-layer neural networks [J]. Microelectronics Journal, 2017, 66: 31-40.

[12] Chua L. Memristor-the missing circuit element [J]. IEEE Transactions on Circuit Theory, 1971, 18(5): 507-519.

[13] Strukov D B, Snider G S, Stewart D R, et al. The missing memristor found [J]. Nature, 2008, 453(7191): 80-3.

[14] Zhang K J, Ren K, Qin X Z, et al. Tunable negative differential resistance and resistive switching properties of amorphous WO devices[J]. IEEE Trans Electron Devices, 2021, 68(8): 3807-3812.

[15] Qi Y F, Shen Z J, Zhao C, et al. Resistive switching behavior of solution-processed AlOx and GO based RRAM at low temperature [J]. Solid-State Electronics, 2020, 168: 107735.

[16] Brivio S, Spiga S, Ielmini D. HfO_2 - based resistive switching memory devices for neuromorphic computing [J]. Neuromorphic Computing and Engineering, 2022, 2(4): 042001.

[17] Song W D, Wang W J, Lee H K, et al. Analog switching characteristics in TiW/Al_2O_3/Ta_2O_5/Ta RRAM devices [J]. Applied Physics Letters, 2019, 115(13): 133501.

[18] Zhang L, Tang Z H, Fang J L, et al. Synaptic and resistive switching behaviors in NiO/Cu_2O heterojunction memristor for bioinspired neuromorphic computing [J]. Applied Surface Science, 2022, 606: 154718.

[19] Jo S H, Chang T, Ebong I, et al. Nanoscale memristor device as synapse in neuromorphic systems [J]. Nano Letters, 2010, 10(4): 1297-301.

[20] Zidan M A, Fahmy H A H, Hussain M M, et al. Memristor-based memory: The sneak paths problem and solutions [J]. Microelectronics J, 2013, 44(2): 176-183.

[21] Yakopcic C, Taha T M, Subramanyam G, et al. Analysis of a memristor based 1T1M crossbar architecture [C]// 2011 International Joint Conference on Neural Networks. IEEE, 2011: 3243-3247.

[22] Zhang Y, Wang X, Friedman E G. Memristor-based circuit design for multilayer neural networks [J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2018, 65(2): 677-686.

[23] Neamen D A. Microelectronics: circuit analysis and design [M]. New York: McGraw-Hill, 2007.