

基于忆阻器边缘计算的图像分类电路设计

罗佳¹, 冉欢欢², 何凯霖¹, 丁晓峰¹

(1. 四川大学锦江学院, 四川眉山 620860; 2. 西华师范大学, 四川南充 637002;)

摘要: 本文针对边缘智能设备低功耗、轻算力的要求, 采用了新型存算一体器件-忆阻器作为基础电路元件, 设计了低功耗、图像识别电路。该电路采用多个忆阻卷积层和忆阻全连接网络串联的方式, 获得了较高的识别精度。为了减小忆阻卷积层计算所需的忆阻交叉阵列的行尺寸和列尺寸的不平衡, 同时降低输入电压方向电路的功耗, 将输入电压反相器置于忆阻交叉阵列之后。该电路可以将完成忆阻卷积网络运算所需的忆阻交叉阵列的行大小从 $2M+1$ 减少到 $M+1$, 同时将单个卷积核计算所需的反相器的数量降低到了 1, 从而大幅度降低了忆阻卷积网络的体积和功耗。利用数学近似, 将 BN 层和 dropout 层计算合并到 CNN 层中, 减小网络层数同时降低电路的功耗。通过在 CIFAR-10 数据集上的实验表明, 该电路可以有效地对图像进行分类, 同时具备推理速度快 (136ns) 和功耗低的优点 (单个神经元功耗小于 $3.5\mu\text{W}$)。

关键词: 边缘计算; 忆阻器; 卷积; 反相器; 数据集; 神经元

中图分类号: V211.3

文献标志码: A

DOI: 10.13195/j.kzyjc.2021.0120

Design of Image Classification Circuit Based on Edge Computing of Memristor

LUO Jia¹, RAN Huan-huan², HE Kai-lin¹, DING Xiao-feng¹

(1. Sichuan University Jinjiang College, Meishan 620860, China; 2. China West Normal University, Nanchong 637002, China)

Abstract: Aiming at the requirements of low power consumption and light computing power for edge smart devices, this paper uses a new type of integrated storage and computing device-memristor as the basic circuit element, and designs low power consumption and image-specific circuits. The circuit uses a series of multiple memristive convolutional layers and a memristive fully connected network to obtain high recognition accuracy. In order to reduce the imbalance of the row size and column size of the memristive interleaved array required for calculation of the memristive convolutional layer, and at the same time reduce the power consumption of the input voltage direction circuit, the input voltage inverter is placed after the memristive interleaved array. This circuit can reduce the row size of the memristive interleaved array required to complete the memristive convolution network operation from $2M+1$ to $M+1$, and at the same time reduce the number of inverters required for the calculation of a single convolution core to 1, This greatly reduces the volume and power consumption of the memristive convolutional network. Using mathematical approximation, the calculations of the BN layer and the dropout layer are merged into the CNN layer to reduce the number of network layers and reduce the power consumption of the circuit. Experiments on the CIFAR-10 data set show that the circuit can effectively classify images, while having the advantages of fast inference speed (136ns) and low power consumption (the power consumption of a single neuron is less than $3.5\mu\text{W}$).

Keywords: Edge computing; memristor; Convolution; Inverter; Data set; Neuron

0 引言

随着近年来人工智能的蓬勃发展, 大量卷积神经网络相继被提出, 而进一步依托于互联网的高速发展, 由大量数字处理设备和服务器所组成的大型数据计算中心相应产生, 这为各类模型的提出和运

算提供了极佳的实验基础, 同时数据计算中心的不断扩大, 也相应促进互联网技术的相应发展, 云计算正式诞生。云计算拥有大量的应用场景和需求, 在商用上, 类似于谷歌和亚马逊这样的互联网公司早早布局 [1-2], 拥有了持续的用户数据接入。云计算目前拥有广阔的市场和用户, 主要通过网络通信接

基金项目: 国家自然科学基金 (项目编号: 51208434); 四川省科技厅科技计划项目 (项目编号: 2021YJ0315).

[†]通讯作者. E-mail: 364483391@qq.com.

入海量的用户数据运算后,反馈给用户。显然,整个过程极度依赖网络通信,对通信要求较高,同时由于双向通信会损失一部分的时间延时,对应高响应要求的处理事件不友好。而涉及到用户数据,对于用户的隐私保护和接入管理也同样重要,云计显得难以处理以上问题。在这样的情况下,边缘计算进入了大家的视野,将用户数据保留于边缘端,减少数据对于云计算的上传,降低通信要求,在边缘端部署部分算力,满足低功耗要求,并且对用户隐私数据进行实时保护[3]。边缘设备的计算方式正是解决上述问题的合理设计。而在对于海量图像数据分析和处理领域,深度学习网络拥有极佳的效率和性能,这可以满足图像识别对效率和精度的要求。而图像数据的分析和处理的效率及性能极大程度取决于神经网络的构架,Iandola等人提出了一个名为“squeezeNet”的小型卷积神经网络架构[4]。Szegedy等人提出了一个更加深入的卷积神经网络架构,称为Inception[5]。Dai等人对Szegedy的研究进行了改进[6]。Zeiler和Fergus提出了一种可视化卷积神经网络激活函数[7],Google的研究人员提出了BlazeFace[8]一个超高性能的人脸检测网络,通过BlazeFace,移动平台上可以达到1ms以内的检测速度,拓宽了很多基于边缘计算的人脸相关联应用的发展空间。Krizhevsky等人提出了一种深度卷积神经网络结构AlexNet,这是深度学习的一个重大突破[9]。AlexNet由5个卷积层和3个全连接层组成。这种架构使用图形处理单元(GPU)进行卷积运算,ReLU作为激活函数,ReLU激活函数不仅在计算方面比Sigmoid更加简单,而且可克服Sigmoid函数在接近0和1时难以训练的问题。并使用dropout来减少过拟合问题[10]。传统深度学习对于计算能力的需求极大,传统的冯诺依曼计算机体系结构通过运算器和控制器进行数据处理,利用提取和存储存储器内的数据,数据传输效率较低,而神经态计算计算结构,有别与传统的冯诺依曼计算机体系结构,以生物神经元系统的信息传递为参照,实现神经系统类似的神经元的计算[11]。而作为模拟神经元的基础电路元件所需的元件信息的传递功能,系统记忆学习能力和高度灵敏性也是目前神经态计算系统面临的主要问题。随着近年来元件领域的发展,纳米级原件出现并迅速被广泛应用,比传统CMOS体积更小的纳米硅薄膜晶体管的出现,不仅大大降低了电路的功耗,电路的速度也得到极大的提升[12]。忆阻器的出现为上述问题提供了理想的解决方案。忆阻器是具有动态

特性的电阻器,根据激励电压其阻值可变[13]。与传统电路元件相比,忆阻器的记忆性和灵活性使其更适应于神经态计算计算结构。并且由于忆阻器纳米级尺寸和高速的特性,非常适合边缘设备的低功耗和速度要求。因此,忆阻器可以作为神经态计算计算结构的基础电路元件进行神经网络突触功能的记忆储存功能,实现深度学习网络的结合,拥有广阔的应用空间[14]。T.A.Bala等人提出针对忆阻器深度学习神经网络的高效激活电路,从电路设计方向对忆阻神经网络进行优化[15]。S.wen和H.Liu和J.Chen等人针对忆阻器神经网络实际应用场景对网络进行裁剪和优化,通过二值化和稀疏化等方法降低网络对忆阻器性能要求,提高性能[16-18]。G.Bao等人针对忆阻器模型与组成电路特性,对比传统网络电路,基于电压电流关系,仿真证明忆阻器网络模型性能和优越性[19]。G.Wan等人通过联系忆阻器再神经网络中的应用,从电路和网络上设计了一种具有改进的忆阻器遗忘模型,通过仿真证明有效性,在模拟人脑运算及记忆方面提供了可能性[20]。通过上述工作,基于忆阻器构建的神经网络电路为边缘设备计算有着广阔的应用前景和大量的实践证明,本文从神经网络各层的运算要求出发,对整体网络构架优化电路,设计出适合各层网络的忆阻器运算电路,依托于忆阻器记忆和计算一体的特性,提高图像识别的效率,降低运算所需的功耗。

1 基本忆阻器神经电路

1.1 忆阻器模型

忆阻器的常用模型采用了其电压阈值的特性,只有当输入电压超过忆阻器的阈值电压时忆阻器的阻抗才会发生变化。输入电压 $V(t)$,忆阻器阻抗 $G(t)$,和当前产生电流 $I(t)$ 有以下关系:

$$V(t) = G(t)I(t) \quad (1)$$

忆阻器阻抗的微分方程可以表示为:

$$\frac{dG(t)}{dt} = -L(v(t))H(G(t), V(t)) \quad (2)$$

$H(G(t), v(t))$ 是一个窗口函数,将忆阻器的阻抗限制在 $[\frac{1}{G_{on}}, \frac{1}{G_{off}}]$ 内。

$$H(G(t), v(t)) = \theta(-v(t))\theta(\frac{1}{G_{off}} - G(t)) + \theta(v(t))\theta(H(t) - \frac{1}{G_{on}}). \quad (3)$$

$\theta(x)$ 是一个参数方程,当 $x > 0$ 时, $\theta(x) = 1$,

当 $x < 0$ 时, $\theta(x) = 0$, $L(v(t))$ 阈值电压函数如下:

$$L(v(t)) = \varphi[V(t) - 0.5(|v(t) + Vt| - |v(t) - Vt|)]. \quad (4)$$

当 $v(t) > V(t)$ 时, $T(v(t)) > 0$, 忆阻器阻抗 $G(t)$ 会发生变化, φ 是忆阻器阻抗变化率。

1.2 忆阻器交叉阵列和记忆卷积层

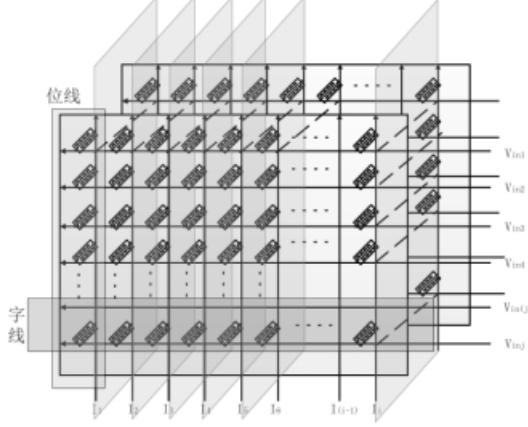


图 1 基于忆阻器 crossbar 的向量和矩阵点乘运算图

V_{in} 是外围控制电路的输入 (j 是忆阻器交叉阵列的字线号数, V_{in} 是一个 j 维向量)。根据基尔霍夫电流定律, 电路中任一个节点上在任一时刻, 流入节点的电流之和等于流出节点的电流之和。

$$L_{out} = V \lambda_{in}^T \times G. \quad (5)$$

I 是一个 N 维向量, G 是一个 $M \times N$ 的电导矩阵。对于一个 $M \times N$ 的忆阻器交叉阵列, $M \times N$ 维加法和乘法可同时进行, 极大地提高了矩阵运算效率并且降低了功耗。而随着忆阻器交叉阵列规模的增加, 计算量需求更大的矩阵和向量也可以得到满足。记忆卷积层的计算主要是向量与矩阵的点乘运算, FC 层可以看作是一种 $H=1, W=1$ 的特殊卷积层。记忆卷积层由多个卷积核组成, 输出特征映射由卷积核的卷积结果和输入数据组成。令第 k 次记忆卷积层的输入图像为 $H \lambda_{in}^k \times W \lambda_{in}^k \times C \lambda_{in}^k$, 卷积核为 $n \times n$, 步长为 1, 其中 n 为用户决定的参数。当使用相同的填充模式时, 输出特征映射为 $H \lambda_{out}^k \times W \lambda_{out}^k \times C \lambda_{out}^k$ 。 (i, j, c) 输出特征可表示为

$$y_{i,j,c}^k = \sum_{m=-\frac{n}{2}}^{\frac{n}{2}} \sum_{n=-\frac{n}{2}}^{\frac{n}{2}} \sum_{k=0}^{C_{in}} I_{in}(i+m, j+n, k) \times W(m, n, k, c). \quad (6)$$

如果 i 和 j 是不变的, 那么第 c 次卷积核 W^c 和输入数据 $I_{in}^{i,j}$ 是不变的, 可以在忆阻器交叉阵列位线中得到一个输出特征映射的值。通过上图所示忆阻器交叉阵列, 我们需要 $(2 \times N^2 + 1) \times C_{in}$ 个忆阻器交叉阵

列才能在一个卷积计算周期内得到一个 $1 \times 1 \times C_{out}$

1.3 忆阻卷积神经网络上的批量标准化

在深度神经网络中批量标准化层是一种典型的处理层。主要是为标准化每一层的输入。理想的数据标准化公式如下:

$$\hat{x}_{bn}^{(k)} = \frac{x_{bn}^{(k)} - E[x_{bn}^{(k)}]}{\sqrt{Var[x_{bn}^{(k)}]}}. \quad (7)$$

数据标准化将影响下一层网络学习提取的特征。因此, BN 层还会包括数据重建功能。BN 层的输出如下:

$$\gamma_{bn}^{(k)} = \beta_{bn}^{(k)} \hat{x}_{bn}^{(k)} + \rho_{bn}^{(k)}, \quad (8)$$

$\beta_{bn}^{(k)}$ 和 $\rho_{bn}^{(k)}$ 是本层的学习参数, 在反向传播网络中, BN 层的参数包括 $E[x_{bn}^{(k)}]$, $D[x_{bn}^{(k)}]$, $E[x_{bn}^{(k)}]$ 和 $D[x_{bn}^{(k)}]$ 是网络训练过程中所有数据的均值和方差的统计值。在后面的网络中这两个参数可以通过下述方程计算:

$$\begin{aligned} \widehat{E[x_{bn}^{(k)}]} &= (1 - momentum) \times E[x_{bn}^{(k)}] + \\ &momentum \times \widehat{E[x_{bn}^{(k)}]}_t. \end{aligned} \quad (9)$$

$$\begin{aligned} \widehat{D[x_{bn}^{(k)}]} &= (1 - momentum) \times D[x_{bn}^{(k)}] + \\ &momentum \times \widehat{D[x_{bn}^{(k)}]}_t. \end{aligned} \quad (10)$$

$\widehat{E[x_{bn}^{(k)}]}$ 和 $\widehat{D[x_{bn}^{(k)}]}$ 是当前输入数据的均值和方差。当 $momentum$ 趋近于 0, 网络训练结束时 $\widehat{E[x_{bn}^{(k)}]}$ 和 $\widehat{D[x_{bn}^{(k)}]}$ 是近似值。因此 BN 层的输出可以简化为输入 X_{bn} 的一个线性变换:

$$Y_{bn} = M_{bn} X_{bn} + B_{bn}. \quad (11)$$

M_{bn} 是对角矩阵。如果在网络中 BN 层之后是 CNN 显然 $X_{bn} = Y_{cnn}$ 在 CNN 中, 输入 X_{cnn} 和输入 Y_{cnn} 的关系可以表示为:

$$Y_{cnn} = M_{cnn} X_{cnn} + B_{cnn}. \quad (12)$$

在下一个网络中设 $momentum=0$, 因此

$$Y_{bn} = M_{bn} M_{cnn} X_{cnn} + B_{bn}. \quad (13)$$

令 $M_{bn} = M_{bn} M_{cnn}, B_{bc} = M_{bn} B_{cnn} B_{bn}$, 则

$$Y_{bc} = M_{bc} X_{cnn} + B_{bc}. \quad (14)$$

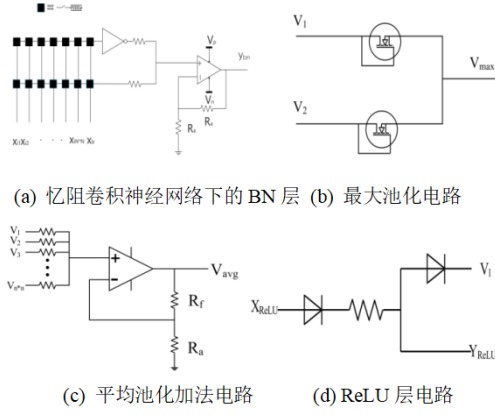


图2 忆阻卷积神经网络下的电路设计

由于式 (7) 和式 (14) 意义相同, 因此可只改变忆阻器交叉阵列的系数, 便可以使 BN 层和 CNN 层共用一个忆阻器交叉阵列。当一个内核为 N 的 CNN 层后连接着一个 BN 层时, 具体电路如图 2(a) 所示, 通过共用忆阻器减少了运算放大器的使用, 这样大大降低了功耗。

1.4 池化操作

池化层将输入的特征映射分割为多个 $N \times N$ 的单元进行最大和平均值的计算, 进一步简化提取特征, 因此提取的特征图像尺寸变为原来的 $\frac{1}{N^2}$ 。最大池化和平均池化是池化层的主要操作方式。最大池化通过并联的两个 nmos 二极管选择输入中的最大电压, 如图 2(b) 所示, 若输入数大于 2, 可以通过多个下述结构组合完成。平均池化一般通过加法电路实现, 如图 2(c) 所示, 平均池化输出一般可以通过下述计算:

$$V_{avg} = \frac{N \times N}{\sum_{i=1}^{N \times N} V_i} (1 + \frac{R_f}{R_a}). \quad (15)$$

当 $R_f \ll R_a$, 可表示为:

$$V_{avg} = \frac{N \times N}{\sum_{i=1}^{N \times N} V_i}. \quad (16)$$

因此可以用这样的加法电路可以用于忆阻神经网络池化层。

1.5 限流二极管下的 ReLU

ReLU 函数是人工神经网络中的典型激活函数 [21]。其计算简单, 有效避免梯度消失和梯度爆炸的特性, 在深度神经网络中得到了广泛的应用。ReLU 函数通过以下方式将线性输入非线性化。

$$F(x) = \max(0, M^T X + B). \quad (17)$$

ReLU 的实现相当于单向限制电路, 只允许电压大于 1 的信号通过, 因此限流二极管非常适合担当 ReLU 电路的基础元件。但由于 ReLU 层的输出可能会大于忆阻器的阈值, 所以在 ReLU 后加上限制电路, 实

现方式如下。

$$F(x) = \min(\max(0, M^T X + B), V_l). \quad (18)$$

在本文实现 ReLU 层选用限流二极管的另一个原因是功耗较低, 在此基础上, 只需加上一个上限二极管, 实现如图 2(d) 所示。当导通时电路存在压降, 所以在理论计算时应该加上二极管的偏置电压。

2 忆阻卷积神经网络的图像分类

2.1 概述

忆阻卷积神经网络由 5 个卷积层和 2 个全连接层组成, 其输入图像是尺寸为 $32 \times 32 \times 3$ 的 RGB 色彩图像。前四个卷积层每层之间由一个最大池化层连接, 第五个卷积层后经过平均池化层输出到全连接层。全连接层输出后通过递增函数 softmax(x) 层将分类网络结果 x 进行标准化映射到 0-1。softmax(x) 层输出值最大的通道代表输入图像的分类结果, 如图 3 所示。

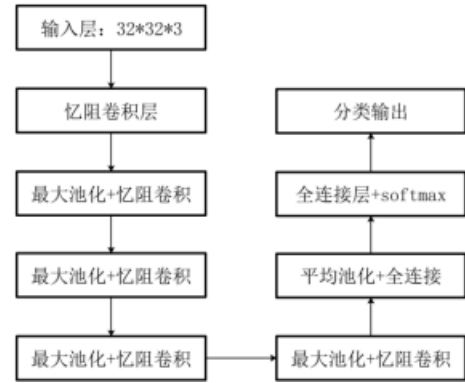


图3 基于忆阻器神经网络的图像分类流程图

2.2 忆阻卷积层

根据文章 [22], 卷积层的参数可以分离为一个正系数矩阵 K^+ 和一个负系数矩阵 K^- , 当 $\sigma_b > 0$, $x_b > 0$ 时

$$V_{o1} = \sum_{i,j}^{N \times N} x_i \sigma_{i,j}^+ + x_b \sigma_b. \quad (19)$$

$$V_{o2} = \sum_{i,j}^{N \times N} x_i \sigma_{i,j}^-. \quad (20)$$

$$\gamma_j = V_{o1} - V_{o2} = \sum_{i,j}^{N \times N} x_i \sigma_{i,j}^+ - x_i \sigma_{i,j}^- + x_b \sigma_b. \quad (21)$$

$\sigma_b < 0$, $x_b > 0$ 时

$$V_{o1} = \sum_{i,j}^{N \times N} x_i \sigma_{i,j}^-. \quad (22)$$

$$V_{o2} = \sum_{i,j}^{N \times N} x_i \sigma_{i,j}^+ - x_b \sigma_b. \quad (23)$$

$$\gamma_j = V_{o1} - V_{o2} = \sum_{i=1}^{N \times N} x_i \sigma_{i,j}^+ - x_i \sigma_{i,j}^- - x_b \sigma_b. \quad (24)$$

对图像分类的第一层忆阻卷积层，卷积核的大小为 $3 \times 3 \times 3$ ，输出通道数为 32。为了实现卷积核的计算需求，则需要忆阻器交叉阵列的规模为 $(3 \times 3 \times 3 + 1) \times 32 \times 2$ 。通过 DAC 将数据转换输入进忆阻器交叉阵列，经过 BN 层和 RELU 处理后，通过 ADC 转换将数据送出，如图 4(a) 所示。

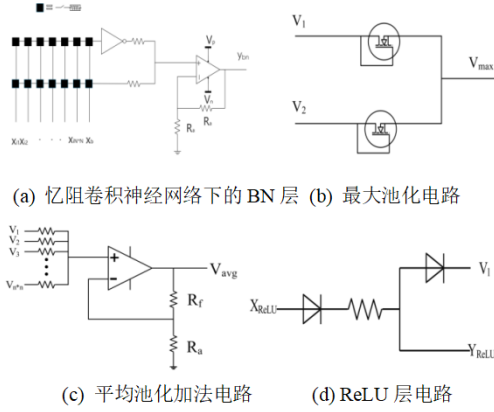


图 4 忆阻卷积层设计

2.3 最大池化和忆阻卷积层

在第一个忆阻卷积层输出后，经过最大池化层处理后的数据输入第二个忆阻卷积层，第二次忆阻卷积的卷积核尺寸为 $3 \times 3 \times 3$ ，输出通道数为 64，为了实现卷积核的计算需求，则需要忆阻器交叉阵列的规模为 $(3 \times 3 \times 32 + 1) \times 64 \times 2$ ，如图 4(b) 所示。

2.4 平均池化和全连接层

在第五个忆阻卷积层输出后，经过平均池化层处理后的数据输入全连接层。将之前忆阻卷积层计算的特征映射为 $3 \times 3 \times 256$ 映射为 $1 \times 1 \times 512$ 的特征向量，需要平均池化电路实现，为了提高计算速度，可以通过同时铺设多个平均池化电路，降低计算周期数来减少电路搭建，在一个周期内结束运算，提高整体运算速率，如图 4(c) 所示。

3 实验分析

3.1 精准度分析

目前国内外学者根据忆阻器的电路特性，建立了大量完善的 MATLAB 的仿真模型，本文的仿真模型采用的是学者们公认的密歇根大学卢伟教授提出的基于 MATLAB 的忆阻器电路模型 [23]，能够有效模拟忆阻器的电学特性。本文首先建立忆阻器的 MATLAB 电路模型，然后对神经网络中的各层进行了电路设计，完成了忆阻卷积神经网络电路的构建。通过文章 [24] 对忆阻器网络参数进行编程。输入尺

寸为 $32 \times 32 \times 3$ 的 RGB 图像，输出对该图像的十种分类的预测。数据上采用 CIFAR-10 数据集，该数据集包含十种分类的对象，每种分类图像数量为 6000 张，总共 60000 张图像数据集。本文实验采用十个分类每个分类 5000 张，总共 50000 张图像进行训练，最后 10000 张图像进行检验测试，准确率大于 %86，如图 5 所示。

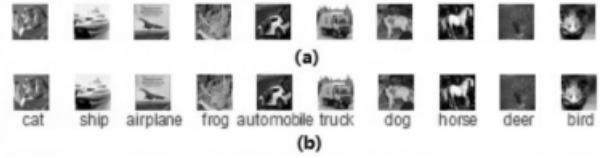


图 5 基于忆阻神经网络的图像识别分类结果

3.2 能耗分析

在忆阻卷积神经网络中有五个忆阻卷积层和两个全连接层。从忆阻卷积神经网络的图像分类中可以得出卷积层的等效电导率分布，如图 6(a) 所示，并由等效电导率可以计算出忆阻卷积神经网络的功耗。功耗主要由两部分组成，一是来源于写入电导率，二是预测过程产生功耗，上述两个过程组成了忆阻卷积神经网络工作的主要流程。基于忆阻卷积神经网络的图像分类功耗可以通过下述公式得到：

$$P = \frac{W}{t} = \frac{(\int_{t_0}^{t_i} U_1^2 G(t) dt + \int_{t_i}^{t_r} U_2^2 G dt)}{t_r - t_0}. \quad (25)$$

上式中，U 是通过单个忆阻器的电压，G 为忆阻器的等效电导率，W 是忆阻器消耗的功率，由于忆阻器的 $\frac{1}{G_{on}}$ 和 $\frac{1}{G_{off}}$ 分别为 $1K\Omega$ 和 $1M\Omega$ ，所以忆阻器的等效电压电导率分别为 10_{-3} 和 10_{-6} 。为了匹配网络的权重，网络权重和忆阻器电导率关系如下：

$$W_{ij} = 1000G_{ij}. \quad (26)$$

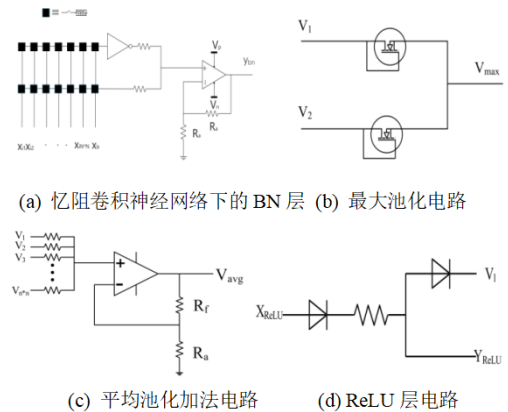


图 6 MCCN 中忆阻器能耗分析结果

令 $\beta=0.6$ ，写入电压为 $G(t)=6.5V$ ，则可得出每个忆阻卷积层的功耗，如图 6(b) 在导通阶段，忆阻卷积

平均功耗为 3.5uW。在网络中的预测过程, 功耗随着输入数据变化而变化, 可以采用写入电压最大时的最大功耗表示, 如图 6(c) 所示。显然, 由以上两个过程组成的忆阻卷积神经网络的主要功耗, 远远小于 CMOS 电路下同样计算过程的功耗 [25]。表一展示了忆阻卷积神经网络所需要的忆阻器资源量, 由于忆阻器的尺寸一般为方形, 为了尽可能发挥忆阻器的优势并且合理利用资源, 通常使忆阻器的行和列数目尽可能接近。表中可看出需要资源量最大的过程为第四个 MCCN 层, 所以这里设计忆阻器交叉阵列规模为 2305×2305 , 电路便可以满足网络要求。忆阻器的响应时间在 100ps 以下, 如果按照每个计算周期 100ps 计算, 则本文设计的忆阻神经网络电路所需的推理时间为: 136.5ns。与文献 [26] 相比, 本文用多个最大池化电路在一个时钟周期内完成计算, 将最大池化层的计算周期从 512 个时钟周期降到了 1 个时钟周期。因此本文设计的电路的推理时间更少, 电路速度更快。在忆阻器的功耗方面, 本文的设计的电路的单个忆阻器的功耗略高于文献 [26] 的数据, 但本文设计的电路只需要 5 个忆阻阵列, 而文献 [26] 需要 17 个忆阻阵列, 在忆阻阵列的数量方面远小于对比文章的结果。忆阻阵列的数量直接决定了忆阻神经网络电路的总功耗和面积。忆阻阵列的数量越多, 忆阻神经网络电路的体积越大, 总功耗越高。因此本文设计的电路在推理

表 1 图像分类中所需忆阻器交叉阵列规模

层	MCs 规模	计算周期
MCNN	28×48	32×32
MCNN	289×128	16×16
MCNN	576×256	8×8
MCNN	1153×512	4×4
MCNN	2305×1024	2×2
GAP	0	1
FC	2×512	1

4 结论

本文基于新型计算存储材料, 忆阻器构建了忆阻神经网络电路用于图像识别, 并成功对 CIFAR-10 图像进行了分类验证。通过实验, 忆阻卷积神经网络能比传统计算构架以更高的实时性和更低的功耗进行图像分类, 通过对各层网络电路的设计和优化, 使计算构架更适合于当前的神经网络和所针对的目标分类。通过优化, 将卷积网络运算所需的忆阻交叉阵列的行大小从 $2M+1$ 减少到 $M+1$, 通过对忆阻器交叉阵列系数的更改, 在 BN 层和 CNN 层共用一个忆阻器交叉阵列, 对运算放大器的替代, 同时在

ReLU 实现电路上通过二极管的限流, 大大降低了功耗。实际电路的特征与目前 MATLAB 的仿真模型存在一定的差异, 我们将在后续的实验中进行进一步对仿真模型进行优化。同时我们也将开展忆阻神经网络电路的设计与验证工作, 来证明本文的仿真电路的有效性。

参考文献 (References)

- [1] M. Bahrami, "Cloud Computing for Emerging Mobile Cloud Apps," 2015 3rd IEEE International Conference on Mobile Cloud Computing, Services, and Engineering, San Francisco, CA, 2015, pp. 4-5, doi:10.1109/MobileCloud.2015.40.
- [2] H. Mueller, S. V. Gogouvtis, H. Haitof, A. Seitz and B. Bruegge, "Poster Abstract: Continuous Computing from Cloud to Edge," 2016 IEEE/ACM Symposium on Edge Computing (SEC), Washington, DC, 2016, pp. 97-98, doi: 10.1109/SEC.2016.26.
- [3] M. Mukherjee, R. Matam, C. X. Mavromoustakis, et al, "Intelligent Edge Computing: Security and Privacy Challenges," in IEEE Communications Magazine, vol. 58, no.9, pp.26-31, September 2020, doi:10.1109/MCOM.001.2000297.
- [4] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and 0.5 mb model size," arXiv preprint arXiv:1602.07360, 2016.
- [5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1-9.
- [6] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in IEEE International Conference on Computer Vision, 2017, pp. 764-773.
- [7] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in European Conference on Computer Vision. Springer, 2014, pp. 818-833.
- [8] V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran, and M. Grundmann, "Blazeface: Sub-millisecond neural face detection on mobile gpus," arXiv preprint arXiv:1907.05047, 2019.
- [9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems, 2012, pp. 1097-1105.
- [10] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from over-fitting," The Journal of Machine Learning Research, vol. 15, no.1, pp.1929-1958, 2014.

- [11] 王洋昊, 刘昌, 黄如, 杨玉超. 神经形态器件研究进展与未来趋势 [J]. 科学通报, 2020, 65(10): 904-915.
- [12] 杨辉, 段书凯, 董哲康, 王丽丹, 胡小方, 尚柳汀. 基于忆阻器-CMOS 的通用逻辑电路及其应用 [J]. 中国科学: 信息科学, 2020, 50(02): 289-302.
- [13] L.O.Chua and S. M. Kang, "Memristive devices and systems," *Proceedings of the IEEE*, vol. 64, no. 2, pp. 209-223, 1976.
- [14] 王春华, 蔺海荣, 孙晶如, 周玲, 周超, 邓全利. 基于忆阻器的混沌、存储器及神经网络电路研究进展 [J]. 电子与信息学报, 2020, 42(04): 795-810.
- [15] S.A.Bala, X. Yang, A. Adeyemo, et al., "A Memristive Activation Circuit for Deep Learning Neural Networks," 2018 8th International Symposium on Embedded Computing and System Design (ISED), Cochin, India, 2018, pp. 1-5, doi: 10.1109/ISED.2018.8704116.
- [16] S. Wen et al., "Memristor-Based Design of Sparse Compact Convolutional Neural Network," in *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 3, pp. 1431 - 1440, July-Sept. 2020, doi: 10.1109/TNSE.2019.2934357.
- [17] H. Liu, S. Sun, J. Liu, Q. Li, J. Diao and Z. Li, "Binary memristive synapse based vector neural network architecture and its application," in *IEEE Transactions on Circuits and Systems II: Express Briefs*, doi: 10.1109/TCSII.2020.3015337.
- [18] J. Chen et al., "An Efficient Memristor-Based Circuit Implementation of Squeeze-and-Excitation Fully Convolutional Neural Networks," in *IEEE Transactions on Neural Networks and Learning Systems*, doi: 10.1109/TNNLS.2020.3044047.
- [19] G. Bao, Y. Zhang and Z. Zeng, "Memory analysis for memristors and memristive recurrent neural networks," in *IEEE/CAA Journal of Automatica Sinica*, vol. 7, no. 1, pp. 96-105, January 2020, doi: 10.1109/JAS.2019.1911828.
- [20] G. Wan, L. Wang, H. Zou and S. Jiang, "A New Model of Associative Memory Neural Network Based on An Improved Memristor," 2020 39th Chinese Control Conference (CCC), Shenyang, China, 2020, pp. 7589-7594, doi: 10.23919/CCC50068.2020.9188654.
- [21] 范丽丽, 赵宏伟, 赵浩宇, 胡黄水, 王振. 基于深度卷积神经网络的目标检测研究综述 [J]. 光学精密工程, 2020, 28(05): 1152-1164.
- [22] A. M. Hassan, H. H. Li, and Y. Chen, "Hardware implementation of echo state networks using memristor double crossbar arrays," in 2017 International Joint Conference on Neural Networks. IEEE, 2017, pp. 2171-2177.
- [23] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, and W. Lu, "Nanoscale memristor device as synapse in neuromorphic systems," *Nano Letters*, vol. 10, no. 4, pp. 1297-1301, 2010.
- [24] C. Yakopcic, M. Z. Alom, and T. M. Taha, "Extremely parallel memristor crossbar architecture for convolutional neural network implementation," in 2017 International Joint Conference on Neural Networks. IEEE, 2017, pp. 1696-1703.
- [25] S. Kim, B. Choi, M. Lim, J. Yoon, J. Lee, H. D. Kim, and S. J. Choi, "Pattern recognition using carbon nanotube synaptic transistors with an adjustable weight update protocol," *ACS Nano*, vol. 11, no. 3, pp. 2814-2822, 2017.
- [26] H. Ran, S. Wen, S. Wang, Y. Cao, and P. Zhou, "Memristor-based Edge Computing of ShuffleNetV2 for Image Classification," *TCAD*, vol. 99, pp. 1-10, 2020.

作者简介

罗佳(1981—), 女, 教授, 硕士, 从事计算机图像识别等研究, E-mail: luojia₄ou@163.com;

冉欢欢(1987—), 女, 硕导, 博士, 从事边缘智能研究, E-mail: 364483391@qq.com;

何凯霖(1982—), 男, 副教授, 硕士, 从事图像处理研究, E-mail: 20788470@qq.com;

丁晓峰(1985—), 男, 副教授, 硕士, 从事图像处理研究, E-mail: 419416215@qq.com.