物理学报 Acta Physica Sinica
Chinese Physical Society
Institute of Physics, CAS

基于忆阻器的脉冲神经网络硬件加速器架构设计

武长春 周莆钧 王俊杰 李国 胡绍刚 于奇 刘洋

Memristor based spiking neural network accelerator architecture

Wu Chang-Chun Zhou Pu-Jun Wang Jun-Jie Li Guo Hu Shao-Gang Yu Qi Liu Yang

引用信息 Citation: Acta Physica Sinica, 71, 148401 (2022) DOI: 10.7498/aps.71.20220098
在线阅读 View online: https://doi.org/10.7498/aps.71.20220098
当期内容 View table of contents: http://wulixb.iphy.ac.cn

专题: 面向类脑计算的物理电子学

基于忆阻器的脉冲神经网络硬件加速器架构设计*

武长春 周莆钧 王俊杰 李国 胡绍刚 于奇 刘洋†

(电子科技大学电子科学与工程学院, 成都 610054)

(2022 年 1 月 13 日收到; 2022 年 5 月 29 日收到修改稿)

脉冲神经网络 (spiking neural network, SNN) 作为第三代神经网络, 其计算效率更高、资源开销更少, 且仿生能力更强, 展示出了对于语音、图像处理的优秀潜能. 传统的脉冲神经网络硬件加速器通常使用加法器模拟神经元对突触权重的累加. 这种设计对于硬件资源消耗较大、神经元/突触集成度不高、加速效果一般. 因此, 本工作开展了对拥有更高集成度、更高计算效率的脉冲神经网络推理加速器的研究. 阻变式存储器 (resistive random access memory, RRAM) 又称忆阻器 (memristor), 作为一种新兴的存储技术, 其阻值随电压变化而变化, 可用于构建 crossbar 架构模拟矩阵运算, 已经在被广泛应用于存算一体 (processing in memory, PIM)、神经网络计算等领域. 因此, 本次工作基于忆阻器阵列, 设计了权值存储矩阵, 并结合外围电路模拟了 LIF (leaky integrate and fire) 神经元计算过程. 之后, 基于 LIF 神经元模型实现了脉冲神经网络硬件推理加速器设计. 该加速器消耗了 0.75k 忆阻器, 集成了 24k 神经元和 192M 突触. 仿真结果显示, 在 50 MHz 的工作频率下, 该加速器通过部署三层的全连接脉冲神经网络对 MNIST (mixed national institute of standards and technology) 数据集进行推理加速, 其最高计算速度可达 148.2 frames/s, 推理准确率为 96.4%.

关键词: 脉冲神经网络, 阻变式存储器, 存内计算, LIF 神经元, 硬件推理加速器

PACS: 84.35.+i, 85.40.–e, 95.75.Mn **DOI:** 10.7498/aps.71.20220098

1 引 言

近年来, 随着深度学习的发展, 神经网络已经得到了广泛的应用. 通过反向传播、梯度下降等方法训练出神经网络可以协助人类完成复杂的工作[1], 有时甚至可以做出优于人类的决策[2]. 传统的人工神经网络 (ANN) 使用了 MP(McCulloch and Pitts) 神经元模型[3], 通过对前一层神经元信号的加权累加和非线性激活输出来模拟神经元的行为, 但是其生物置信度差、计算效率低、硬件资源开销大. 为了构建一种具有更高生物置信度的神经元模型, 1952 年, Hodgkin 和 Huxley[4] 基于神经元细胞膜电位生理现象的非线性微分方程, 提出 Hodgkin-Huxley(HH) 模型. HH 模型可以更加精确地模拟生物神经元膜电压变化和脉冲发放, 但是由于其存在大量的微分、积分操作, 硬件实现难度大. 相比于 HH 模型, LIF(leaky integrate and fire) 神经元模型[5] 在保留一定程度仿生能力的同时, 极大地降低了计算复杂度, 使之可以轻易部署到硬件平台上, 具有更强的硬件友好性. 基于 LIF 神经元模型的 SNN[6,7] 作为第三代神经网络, 相比于传统的单层感知器和多层感知器具有更强的生物合理性及处理时空信息时潜在的更优计算效率, 同时具有相对较低的硬件设计资源开销.

近年来, 基于 SNN 的神经形态计算[8,9] 受到了广泛关注并得到极大发展. 神经形态计算的一个重要目标是通过硬件模拟生物神经网络计算. 基于冯·诺依曼架构的硬件计算平台 (CPU, GPU 等) 是实现神经形态计算的重要途径之一. 但是冯·诺依曼架构存在存算分离导致的冯·诺依曼瓶颈[10,11], 这使得其在实现神经形态计算时效率低下, 且拥有

* 国家自然科学基金 (批准号: 92064004) 资助的课题.

† 通信作者. E-mail: yliu1975@uestc.edu.cn

极为庞大的功耗、面积开销. ASIC 设计是实现神经形态计算的另一个重要途径. 相比于通用计算平台, ASIC 设计具有高的专用性, 其计算效率更高[12]、功耗更低、硬件资源开销更少. 相比于基于模拟 ASIC 设计的类脑计算芯片[13], 基于数字 ASIC 设计的类脑计算芯片[14–17]可以实现更高的集成度, 因此受到了更加广泛的关注. 为了解决冯·诺伊曼瓶颈对计算效率的影响, 数字 ASIC 设计中往往使用寄存器作为存算单元, 实现存算一体的神经形态芯片设计. 但是随着设计规模的不断增加, 寄存器阵列及其外围电路的复杂度不断上升、资源开销迅速增长.

近年来, 基于忆阻器的存算一体技术[18]由于其资源开销低、计算效率高等特点受到了广泛的关注. 相比于数字集成电路中基于寄存器的存算一体架构, 基于忆阻器的 crossbar 架构复杂度更低, 且集成度更高. 同时, 基于忆阻器 crossbar 架构的类脑计算[19–22]可以并行输入多组脉冲, 通过对输出电流的统计、求和, 从而实现对多路突触的并行计算. 但是, 其输出往往是电流形式, 需要外加运算放大器和 ADC 才能实现模拟电流到数字电压的转换, 其外围电路复杂度高且不易集成. 2020 年, 国防科技大学忆阻器研究团队[23]提出了使用 N 位等值逻辑比较电路实现忆阻器阵列的数字电压输出, 其在一定程度上降低了外围电路的复杂程度, 但是需要更多的时间步长进行计算. 此外, 随着突触集成度的增加, 忆阻器阵列规模需要不断扩大, 资源开销随之增加. 通过突触复用的方式可以有效地解决资源开销随设计规模增扩大而增加的问题, 但是这需要对忆阻器反复擦写来更新突触权值. 尽管这种设计方法降低了设计资源开销, 但是其牺牲了较多的计算效率.

在本次工作中, 致力于解决硬件资源开销与计算效率之间的矛盾, 采用硬件设计语言 (Verilog) 构建了基于忆阻器的 crossbar 阵列作为类脑计算的权值存储阵列. 基于 crossbar 阵列设计, 提出了基于权值共享技术的紧凑化权值存储阵列设计, 并通过神经元、突触复用技术减少忆阻器阵列的擦写频率, 仅使用 0.25k 忆阻器单元模拟 8k 神经元和 64M 突触计算. 此外, 通过在忆阻器阵列的输出端进行串联电阻分压实现了忆阻器阵列输出从电流到电压的转换. 忆阻器阵列的输出结果可以直接通过数字电路采集. 基于权值存储阵列的设计, 构建了紧凑化计算核心. 随后, 基于紧凑化计算核心设计了推理加速器, 并使用通用验证方法学 (UVM) 搭建验证平台. 通过 GPU 对 784×1024×1024×10 规模的全连接网络进行了训练, 并将其部署到加速器平台对 MNIST 数据集进行推理验证实验. 实验结果表明, 该加速器可以实现 148.2 frams/s 的图像识别速率, 同时识别精度达到了 96.4%. 最后, 基于硬件加速设计部署了不同规模的网络进行推理测试, 总结了该设计对于不同规模网络的加速效果并对结果进行了分析.

2 脉冲神经网络相关架构介绍

2.1 脉冲神经网络 LIF 模型结构介绍

LIF 神经元模型是实现脉冲神经网络的一个重要神经元模型. 其相比于传统的 MP 神经元模型, LIF 神经元模型可以通过更少硬件资源来实现, 并且可以更好地模拟生物神经元工作原理. (1) 式所示为 LIF 神经元膜电位与输入脉冲的关系:

$$V_{\mathrm{mp}(t)} = \begin{cases} V_{\mathrm{mp}(t-1)} + \sum_{i=0}^{N-1} O_i \times W_i + V_{\mathrm{L}}, & (|V_{\mathrm{mp}}| < V_{\mathrm{th}}), \\ V_{\mathrm{reset}}, & \text{otherwise}, \end{cases} \tag{1}$$

式中, V_{mp} 是神经元膜电位; O_i 是输入脉冲; W_i 是与输入脉冲对应的权重; V_{L} 是神经元的泄漏电压; V_{th} 是膜电位的阈值; V_{reset} 是神经元膜电位的复位值; N 是神经元个数. 同时将膜电位恢复至, 其膜电位升高. 在一个“time step”的计算中, 神经元完成所有突触的计算, 其膜电位将与阈值电压作对比. 当神经元的膜电位达到阈值电压 (V_{th}) 时, 其将发放一个脉冲, 同时将膜电位恢复至静息值 (V_{reset}); 否则, 神经元将不会发放脉冲, 并将当前膜电位与泄漏电压 V_{L} 相加并保存, 用于下一个“time step”计算.

基于 LIF 神经元模型的四层全连 SNN 结构如图 1 所示. 该网络由一个输入层、两个隐藏层和一个输出层组成, 输入层与各隐藏层分别由 5 个神经元组成, 输出层由两个神经元组成. 相比于传统的 ANN, 该 SNN 网络中神经元的输入和输出都是脉冲形式的, 即 0/1 脉冲电平. 输出端统计两个输出神经元的脉冲发放数量, 经过 n 个 time step”的脉冲统计后, 根据脉冲统计结果进行分类. 脉冲

发放最多的神经元表示其激活程度最高, 即意味着输入的脉冲序列最有可能对应该神经元所代表的类别.

2.2 LIF 神经元的硬件设计方法

一个典型的 LIF 神经元模型的硬件结构如图 2(a) 所示. 其主要实现了脉冲输入、膜电位累加、膜电位比较及脉冲发放. 输入脉冲与对应权值相乘并类形成膜电位, 在膜电位计算完毕后, 神经元通过将之与阈值电压相比较判断是否发放脉冲输出. 其在硬件中的计算原理图如图 2(b) 所示. 输入的脉冲序列由 1/0 脉冲串组成. 输入脉冲通过多路选择器实现与权值的“乘”操作, 选择结果作为突触的输入. 神经元将初始膜电位 (即上一个 “time step”结束后保存的膜电位) 与所有突触的输入相累加, 形成总的膜电位, 并与阈值电压相比较, 如果膜电位超出阈值电压, 则将膜电位清零并保存用作下一个“time step”的初始膜电位, 同时发放一个脉冲输出; 反之, 膜电位将加上一个泄露电压并保存, 同时不产生脉冲输出.

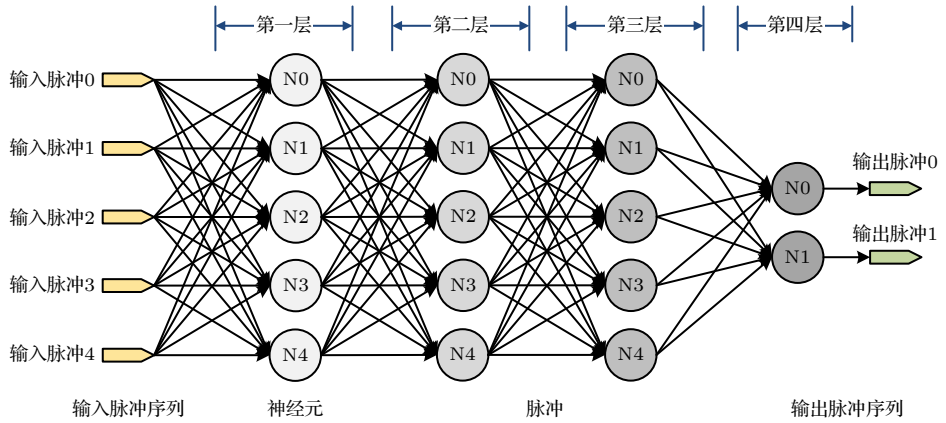

图 1 基于 LIF 模型的全连接脉冲神经网络结构图

Fig. 1. Structure diagram of fully connected spiking neural network based on LIF model.

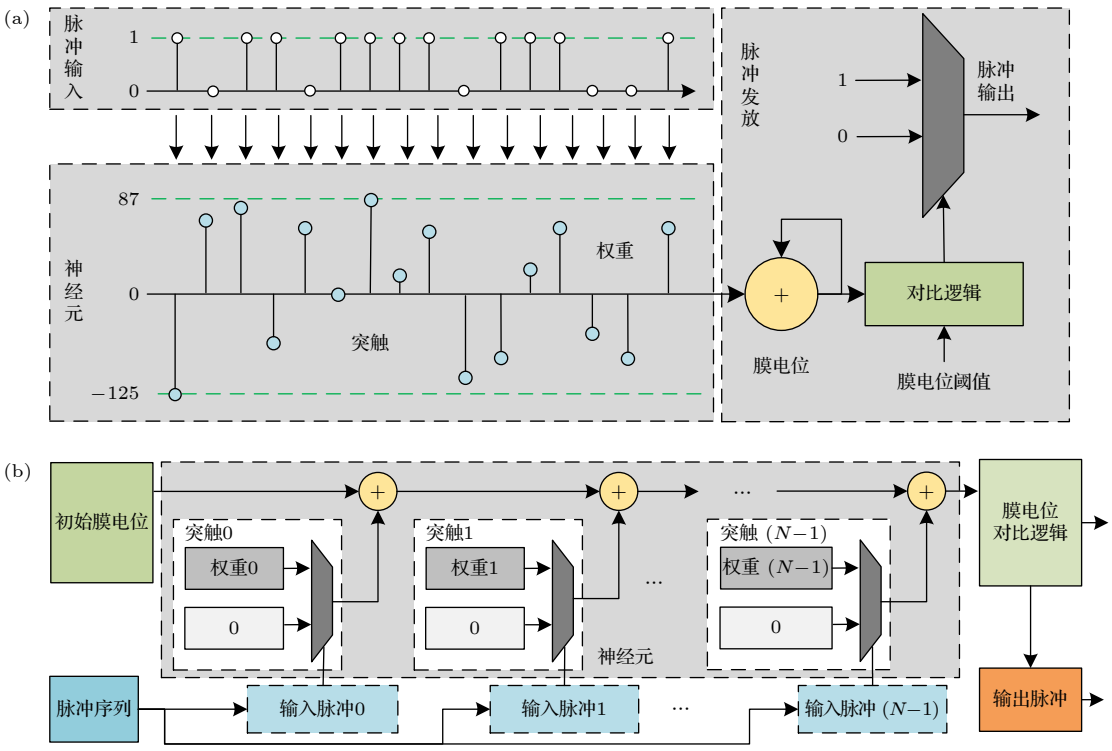

图 2 (a) 神经元原理图; (b) 神经元计算原理图

Fig. 2. (a) Neuron schematic diagram; (b) schematic diagram of neuron computation.

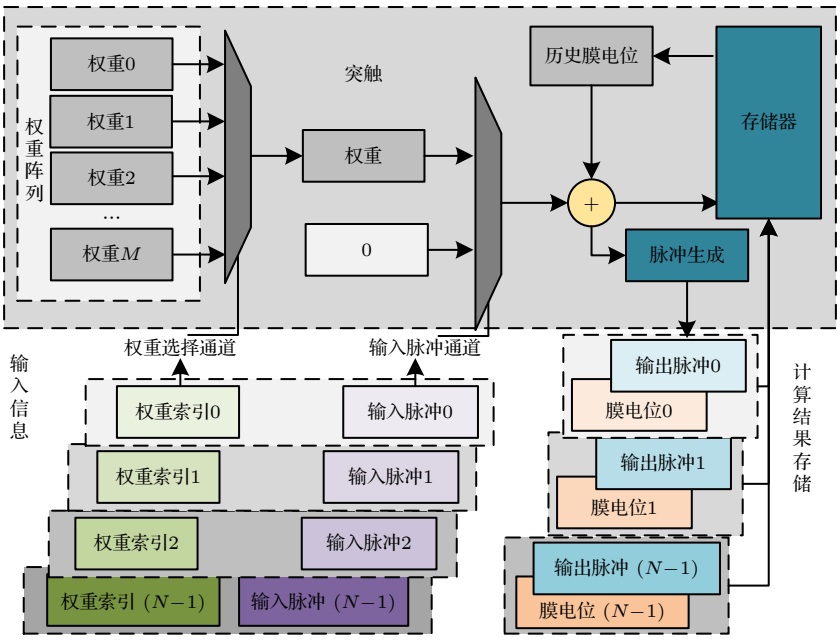

图 3 基于突触复用技术及权值共享技术的 LIF 神经元模型架构

Fig. 3. The LIF neuron model architecture based on synapse multiplexing technology and weight sharing technology.

2.3 基于突触复用技术及权值共享技术的 LIF 神经元模型硬件设计方法

图 3 所示为基于突触复用技术及权值共享技术的 LIF 神经元模型架构. 与传统 LIF 神经元架构不同, 其只有一个脉冲输入通道, 通过分时复用的技术实现所有突触的计算. 所有突触共享 M 个权值, 因此, 每个突触的脉冲输入同时伴随一组权值索引输入, 通过对应的权值索引来获取其对应的权值数据. 当神经元依次完成对所有突触的膜电位累加后, 神经元会对膜电位进行判断并控制脉冲发放.

3 基于忆阻器阵列的脉冲神经网络加速器设计

3.1 基于忆阻器阵列的权重矩阵设计方法

忆阻器由上下两块电极及其中间起隔离作用的氧化层组成, 如图 4(a) 所示. 当其两端加正向电压, 氧化层中会出现氧空位, 从而组成导电细丝, 使得忆阻器导电能力增强, 呈现低阻特性, 其阻值为 $R_{\rm L}$. 当其两端加反向电压, 导电细丝会断裂, 忆阻器呈现高阻特性, 其阻值为 $R_{\rm H}$. 本次工作主要利用了忆阻器的这两种工作状态, 将其作为二值器件使用, 其中忆阻器的高阻与低阻的阻值相差千倍以上.

基于忆阻器单元 A, 设计了 crossbar 结构, 如图 4(b) 所示. crossbar 的行数代表了权值的数量, 每一行代表了一个权值, 每一个忆阻器代表权值的一位, 其高阻态代表逻辑 0, 低阻态代表逻辑 1. 列数则代表了权值的位宽. 当输入端有高电平脉冲输入时, 其对应的权值被激活, 并以电流形式从输出端输出. 每一列的输出端串联一个分压电阻 $R_{\rm D}$, 从而将输出电流转化成电压数据, 其中串联电阻的阻值 $R_{\rm D}=\sqrt{R_{\rm L}R_{\rm H}}$. 因为 $R_{\rm D}$ 远大于忆阻器的低阻阻值且远小于忆阻器的高阻阻值, 因此, 当 $R_{\rm D}$ 对应列被激活的忆阻器阻值为 $R_{\rm L}$ 时, 即可认为输出端是逻辑电平“1”; 当 $R_{\rm D}$ 对应列被激活的忆阻器阻值为 $R_{\rm H}$ 时, 即可认为输出端是逻辑电平“0”. 用这样的方式即可通过数字电路直接对 crossbar 阵列的输出进行结果采集, 而不需要使用运算放大器、数模转换器等模拟电路, 降低了其外围电路的复杂度, 同时提升了计算效率.

3.2 权值共享技术

在神经网络结构中, 一个神经元需要与多个突触相连接, 每个突触都对应有一个权值用来表示连接关系的强弱. 随着网络规模的增加, 突触的数量急剧增加, 忆阻器阵列的资源开销随之增加. 为了减少权值数量从而降低硬件资源开销, 本工作提出了一种基于忆阻器阵列的权值共享技术. 一层网络中所有突触的权值被聚类成 16 个 16-bit 权值 W_i, 有 $\{W_i \in W_0, W_1 \cdots W_{15}\}$. 突触通过一个 4-bit 的索引 S_j 对权值进行索引, 有 $\{S_i \in S_0, S_1 \cdots S_{15}\}$.

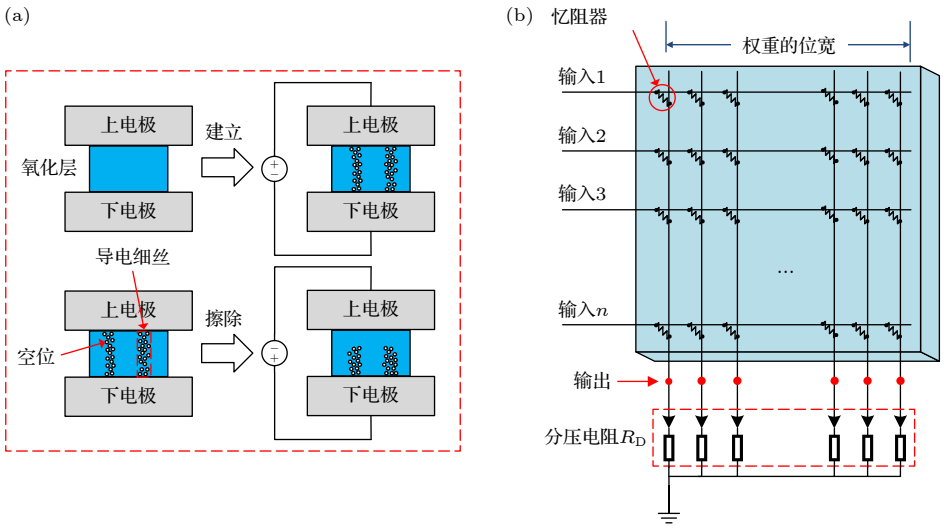

图 4 (a) 忆阻器的建立/擦除示意图; (b) 基于忆阻器的 crossbar 阵列

Fig. 4. (a) The set/reset operation of the memristor; (b) the crossbar structure based on the memristor.

本次工作基于 784 × 1024 × 1024 × 10 的全连接网络开展了对权值共享技术的应用与分析. 实验结果如图 5 所示, 相比于聚类前, 聚类后的权值占用的存储开销从 28.4 Mb (784 × 1024 × 16 bit + 1024 × 1024 × 16 bit + 1024 × 10 × 16 bit) 下降至 0.75 kb, 但网络推理精度仅下降了约 1%. 同时, 由于聚类后权值精度固定为 16-bit, 软件到硬件的映射过程不存在截断误差或舍入误差造成的精度损失, 因此, 权值聚类后的网络可以以同样的精度从软件平台映射到硬件平台.

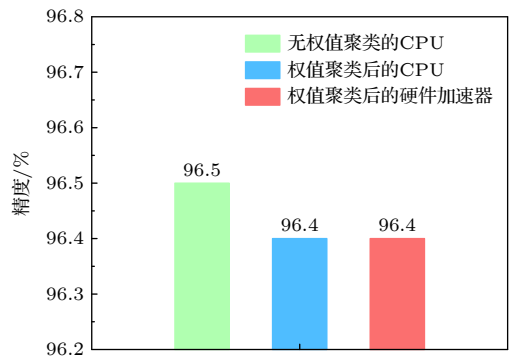

图 5 权值共享技术对精度的影响

Fig. 5. The influence of weight sharing technology on accuracy.

权值量化、权值共享技术通过牺牲很少的网络精度大幅度降低了权值数量与硬件平台资源开销, 是一种对硬件设计十分友好的技术. 通过分析该技术与网络精度的变化发现, 聚类使得突触的权值向不同方向偏移产生 $\pm\Delta W$, 但是宏观上权值偏移量存在互补, 因此网络精度变化不明显. 从另一方面来讲, 神经网络的连接关系相比于权值大小更为关键, 在一定的网络架构下轻微调整权值大小不会对网络精度造成较大影响.

3.3 基于忆阻器阵列的计算核设计

基于忆阻器阵列的计算核架构如图 6 所示, 其中包含一个脉冲整形单元、一个多位的 D 触发器、一个由忆阻器构成的权重矩阵、多位行波计数器和一个忆阻器控制器. 脉冲整形单元实现了对输入脉冲进行计数, 并依据每个脉冲对应的索引值, 将输入脉冲序列转换成 16 个通道的脉冲数据流, 每个通道内的脉冲对应相同的索引. 16 个通道中的脉冲数据通过 D 触发器传递给权重矩阵进行计算. 多位行波计数器统计每个突触的权重, 并进行膜电位累加. 忆阻器控制器可以对忆阻器的阻值进行擦写, 同时也可以将每个“time step”计算后的膜电位与阈值电压进行对比, 从而控制输出脉冲的发放.

在一次计算过程中, 脉冲整形器首先对输入脉冲计数, 并将其整形并预存到不同通道对应的缓存区间中. 每一个通道的脉冲依次送入到权值矩阵中进行权值索引, 同时多位行波计数器对权值进行累加. 直至所有脉冲输入完毕, 忆阻器控制器会将行波计数器的加结果与上一个“time sep”保存的膜电位相加, 并与阈值电压相对比, 控制脉冲发放. 如果膜电位超过阈值电压, 忆阻器控制器会发放一个脉冲输出, 同时清零并保存膜电位; 如果膜电位小于等于阈值电压, 其会保存当前膜电位用于下一个“time step”的计算, 且不会发放脉冲.

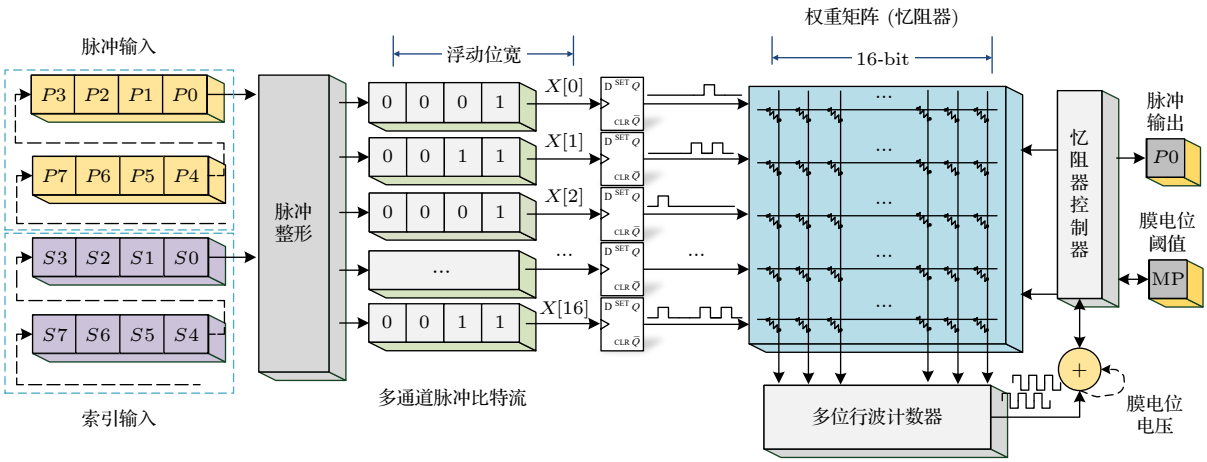

图 6 基于忆阻器阵列的计算核架构

Fig. 6. Computing core architecture based on resistive random access memory matrix.

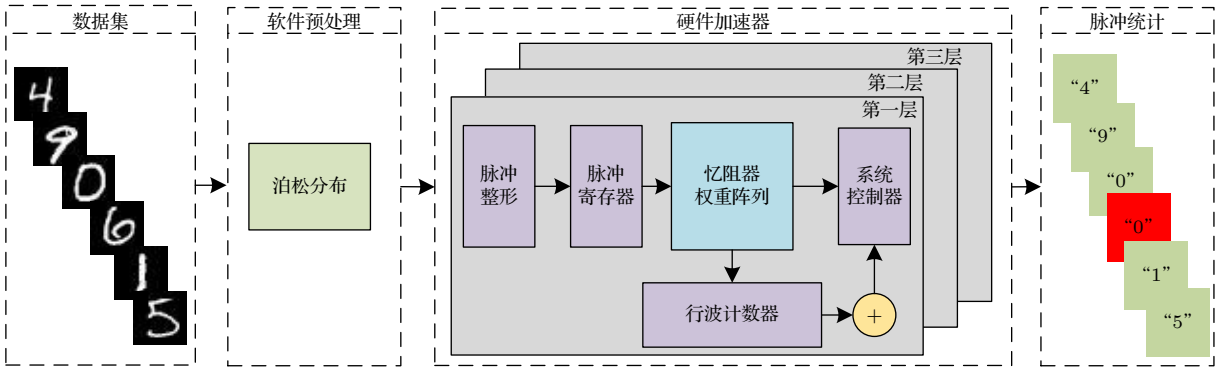

图 7 基于硬件加速器的应用架构图

Fig. 7. Application architecture diagram based on hardware accelerator.

3.4 整体结构

基于计算核的设计, 构建了具有三个计算核的硬件加速器架构, 并设计了完整的应用架构, 如图 7 所示. 软件层通过泊松编码将原始 MNIST 数据集的帧图像转换成脉冲图像, 并送入硬件加速器进行计算. 硬件加速器包含三个前文所述计算核心, 可以同时进行三层网络计算. 硬件加速器在计算完成后, 其脉冲输出被采集、统计并分析, 最终得出计算结果.

3.5 推理网络模型到硬件加速的映射

推理网络模型到硬件加速的映射关系示意图如图 8 所示. 示例中, 输入的帧图像为“9”, 其通过编码转换成 N 个“time step”的脉冲图像, 并输入到硬件加速器中. 硬件加速器中的三个计算核组成了一个三层的计算网络, 第一层接受输入的图像脉冲数据并启动计算, 最后一层将输出脉冲作为计算结果输出至片外. 软件层统计硬件加速器的输出脉冲数, 当 N 个“time step”的脉冲数据计算完毕后, 软件层将输出脉冲数量做对比, 选择累计发放脉冲最多的神经元所对应的标签为本次输入对应的计算结果. 图 8 中, 输出层对应标签为“9”的神经元累计发放的脉冲数量最多, 因此, 软件层将本次计算结果统计为“9”, 与输入脉冲图像标签相符, 是一次正确的计算.

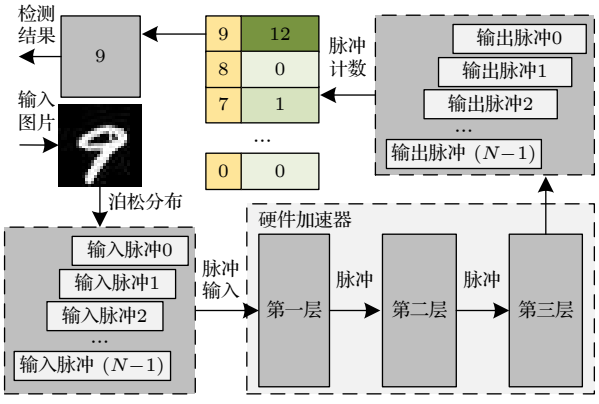

图 8 推理网络模型到硬件加速的映射示意图

Fig. 8. Mapping diagram of network to hardware accelerator.

4 仿真结果及分析

为验证加速器的效果, 采用 Verilog 对模拟忆阻器阵列设计, 并实现整体硬件加速器平台设计, 通过 SPICE 仿真获取内部延时; 采用 UVM 搭建验证平台; 采用 Synopsis 公司的 VCS2018 软件进行编译、仿真, 并对 MNIST 数据集中的 10000 张图片进行识别.

4.1 识别精度

在软件层训练了 784×1024×1024×10 的全连接 ANN 网络, 并通过脉冲转换的方法, 将其转换成 SNN 网络. 软件仿真表明, 其在测试集推理精度为 96.5%. 之后, 在软件层通过 3.2 节所述的权值量化的方法, 分别将三层网络的突触权值量化成 16 个 16-bit 的权值数据, 同时为每个突触生成一个 4-bit 的索引数据. 忆阻器控制器通过对忆阻器阵列的扫描, 将权值数据写入到忆阻器阵列中. 阈值电压、泄漏电压等配置参数被配置到忆阻器控制器中. 软件仿真表明, 权值量化后的 SNN 网络在测试集的推理精度为 96.4%, 相比于量化前, 精度下降了 1%, 但是存储开销降低了约 99.997%.

4.2 识别速度

为验证加速器对不同规模网络的加速效果, 在不考虑精度的前提下, 本工作分别将 4 种不同规模的网络部署到硬件加速器平台, 对 MNIST 数据集进行推理测试. 4 种不同的网络规模分别为 784 × 1024 × 1024 × 10, 784 × 1024 × 2048 × 10, 784 × 1024 × 4096 × 10 和 784 × 1024 × 8192 × 10.

测试结果如图 9 所示. 实验结果表明, 加速器对于 784 × 1024 × 1024 × 10 的网络计算速率为 148.2 frames/s, 随着网络规模的增长, 其计算速率呈线性下降. 通过理论分析发现, 由于硬件加速器的三个计算核心通过并行计算对网络推理进行加速, 因此加速器对网络的加速效率取决于规模最大的一层网络, 即计算量最大的一层网络. 硬件加速器的加速效率随着神经网络中最大规模的一层网络的规模增加而线性下降.

此外, 该加速器对小规模网络表现出了优秀的加速效果, 而对大规模网络的加速效果不明显. 因此, 该硬件加速器架构更适用于计算量相对较小、资源相对匮乏的边缘计算.

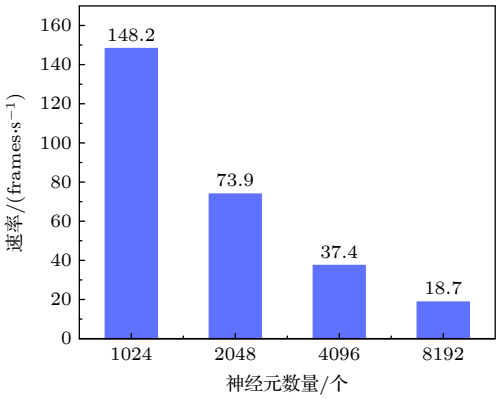

图 9 硬件加速器对不同规模神经网络计算的加速效率

Fig. 9. Acceleration efficiency of hardware accelerator for different scale neuron networks.

4.3 与其他类似工作的对比

表 1 所列为本次工作与近年来其类似工作的对比. 从表 1 中可以看出, 本次工作相较于其他工作, 使用忆阻器实现了更为简单的 LIF 神经元模型, 通过更少的忆阻器资源开销 (0.75k) 集成了更大规模的神经元/突触. 同时, 在 MNIST 数据集的识别工作中拥有相对较高的推理准确率 (96.4%).

表 1 本次工作与其他类似工作的对比

Table 1. Comparison of this work with other works.

	Burr et al.[21]	Peng et al.[24]	Huang et al.[25]	This work
神经元模型	MP	MP	MP	LIF
突触数	161k	1k	~4.8M	192M
神经元数	385	128	~6.4k	24k
忆阻器开销	330k	1k	1.11M	0.75k
测试数据集	MNIST	—	MNIST	MNIST
准确率/%	94.0—97.0	91.7	93.4	96.4

5 总结与展望

在本次工作中, 对基于 LIF 神经元模型的脉冲神经网络进行了详细阐述, 并基于忆阻器阵列及其外围电路实现了 LIF 神经元的硬件设计. 基于神经元/突触复用技术和权值共享技术的权值存储阵列可以通过 0.25k 的忆阻器存储 64M 突触的权值. 此外, 通过电阻分压的方式对忆阻器阵列输出进行电流-电压转换, 并通过数字电路进行采集. 基于忆阻器阵列设计, 提出了神经形态计算核设计. 计算核可以实现至多 8k 神经元、64M 突触的计算. 最终, 通过三个神经形态计算核构建了脉冲神经网

络推理加速器设计, 并开展了仿真、验证工作. 在加速器中部署了 784 × 1024 × 1024 × 10的三层全连接脉冲神经网络, 并对 MNIST 数据集进行了推理验证. 试验结果表明, 该加速器设计可以实现 148.2 frames/s 的推理计算, 同时拥有 96.4%的推理准确率.

受限于忆阻器存储材料的工作频率, 该架构的硬件加速器最高工作频率为 50 MHz, 未来随着工艺的发展, 忆阻器的工作主频将进一步提升, 硬件加速器的性能也能随之得以提升. 此外, 随着部署在加速器中的网络规模不断提升, 加速器的计算效率逐渐下降, 因此, 该加速器设计适合部署在资源匮乏、算力需求低的边缘设备中进行加速工作. 在算力需求较高的应用环境下, 则需要扩大设计规模, 这也就意味着需要更大规模的忆阻器阵列和外围电路.

参考文献

[1] Redmon J, Farhadi A 2017 *30th IEEE Conference on Computer Vision & Pattern Recognition* Honolulu, HI, July 21–26, 2017 pp6517–6525

[2] Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, Hubert T, Baker L, Lai M, Bolton A, Chen Y, Lillicrap T, Hui F, Sifre L, Van Den Driessche G, Graepel T, Hassabis D 2017 *Nature* **550** 354

[3] McCulloch W S, Pitts W 1943 *Bull. Math. Biophys.* **5** 115

[4] Hodgkin A L, Huxley A F 1952 *J. Physiol.* **116** 449

[5] Gerstner W 1995 *Phys. Rev. E:Stat. Phys. Plasmas Fluids Relat. Interdisciplin. Top.* **51** 738

[6] Maass W 1997 *Neural Networks* **10** 1659

[7] Roy K, Jaiswal A, Panda P 2019 *Nature* **575** 607

[8] Chen Y R, Li H, Chen Y Z, Chen F, Li S C, Liu C C, Wen W J, Wu C P, Yan B N 2018 *Artif. Intell. View* **13** 46 (in Chinese) [陈怡然, 李海, 陈逸中, 陈凡, 李思成, 刘晨晨, 闻武杰, 吴春鹏, 燕博南 2018 人工智能 **13** 46]

[9] Schuman C D, Potok T E, Patton R M, Birdwell J D, Dean M E, Rose G S, Plank J S2017 arXiv:1705.06963

[10] Mahapatra N R, Venkatrao B 1999 *Crossroads* **5** 2

[11] von Neumann J 1993 *IEEE Ann. Hist. Comput.* **15** 27

[12] Chen T, Du Z, Sun N, Wang J, Wu C, Chen Y, Temam O 2014 *Acm Sigplan Notices* **49** 269

[13] Benjamin B V, Gao P, Mcquinn E, Chou D Hary S, Chandrasekaran A R, Bussat J, Alvarez-Icaza R, Arthur J V, Merolla P A, Boahen K 2014 *Proc. IEEE* **102** 699

[14] Pei J, Deng L, Song S, Zhao M G, Zhang Y H, Wu S, Wang G R, Zou Z, Wu Z Z, He W, Chen F, Deng N, Wu S, Wang Y, Wu Y J, Yang Z Y, Ma C, Li G Q, Han W T, Li H L, Wu H Q, Zhao R, Xie Y, Shi L P 2019 *Nature* **572** 106

[15] Davies M, Srinivasa N, Lin T H, Chinya G, Cao Y, Choday S H, Dimou G, Joshi P, Imam N, Jain S 2018 *IEEE Micro* **38** 82

[16] Akopyan F, Sawada J, Cassidy A, Alvarez-Icaza R, Arthur J, Merolla P, Imam N, Nakamura Y, Datta P, Nam G J 2015 *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.* **34** 1537

[17] Furber S B, Galluppi F, Temple S, Plana L A 2014 *Proc. IEEE* **102** 652

[18] Li K, Cao R R, Sun Y, Liu S, Li Q J, Xu H 2019 *Micro/nano Electron. Intell. Manuf.* **1** 87 (in Chinese) [李锟, 曹荣荣, 孙毅, 刘森, 李清江, 徐晖 2019 微纳电子与智能制造 **1** 87]

[19] Xia Q F, Yang J J 2019 *Nat. Mater.* **18** 309

[20] Deng Y B, Wang Z W, Zhao C H, Li L, He S, Li Q H, Shuai J W, Guo D H 2021 *Appl. Res. Comput.* **38** 2241 (in Chinese) [邓亚彬, 王志伟, 赵晨晖, 李琳, 贺珊, 李秋红, 帅建伟, 郭东辉 2021 计算机应用研究 **38** 2241]

[21] Burr G W, Shelby R M, Sidler S, Nolfo C D, Jang J, Boybat I, Shenoy R S, Narayanan P, Virwani K, Giacometti E U 2015 *IEEE Trans. Electron Devices* **62** 3498

[22] Moro F, Hardy M, Fain B, Dalgaty T, Clemencon P, De Pra A, Esmanhotto E, Castellani N, Blard F, Gardien F, Mesquida T, Rummens F, Eseni D, Casas J, Indiveri G, Payvand M, Vianello E 2022 *Nat. Commun.* **13** 3506

[23] Fang X D, Wu J J 2020 *Comput. Eng. Sci.* **42** 1929 (in Chinese) [方旭东, 吴俊杰 2020 计算机工程与科学 **42** 1929]

[24] Peng Y, Wu H, Gao B, Eryilmaz S B, Qian H 2017 *Nat. Commun.* **8** 15199

[25] Huang L, Diao J T, Nie H S, Wang W, Li Z W, Li Q J, Liu H J 2021 *Front. Neurosci.* **15** 639526

SPECIAL TOPIC—Physical electronics for brain-inspired computing

Memristor based spiking neural network accelerator architecture*

Wu Chang-Chun Zhou Pu-Jun Wang Jun-Jie Li Guo
Hu Shao-Gang Yu Qi Liu Yang †

(*School of Electronic Science and Engineering, University of Electronic Science and Technology of China, Chengdu* 610054, *China*)

(Received 13 January 2022; revised manuscript received 29 May 2022)

Abstract

Spiking neural network (SNN) as the third-generation artificial neural network, has higher computational efficiency, lower resource overhead and higher biological rationality. It shows greater potential applications in audio and image processing. With the traditional method, the adder is used to add the membrane potential, which has low efficiency, high resource overhead and low level of integration. In this work, we propose a spiking neural network inference accelerator with higher integration and computational efficiency. Resistive random access memory (RRAM or memristor) is an emerging storage technology, in which resistance varies with voltage. It can be used to build a crossbar architecture to simulate matrix computing, and it has been widely used in processing in memory (PIM), neural network computing, and other fields. In this work, we design a weight storage matrix and peripheral circuit to simulate the leaky integrate and fire (LIF) neuron based on the memristor array. And we propose an SNN hardware inference accelerator, which integrates 24k neurons and 192M synapses with 0.75k memristor. We deploy a three-layer fully connected network on the accelerator and use it to execute the inference task of the MNIST dataset. The result shows that the accelerator can achieve 148.2 frames/s and 96.4% accuracy at a frequency of 50 MHz.

Keywords: spiking neural networks, resistive random access memory, processing in memory, leaky integrate and fire model, hardware inference accelerator

PACS: 84.35.+i, 85.40.–e, 95.75.Mn **DOI:** 10.7498/aps.71.20220098

* Project supported by the National Natural Science Foundation of China (Grant No. 92064004).

† Corresponding author. E-mail: yliu1975@uestc.edu.cn