

引用格式:汪晶,王君鹏,孙文昊,等. 用于脉冲卷积神经网络的神经形态处理 VLSI 架构设计[J]. 微电子学与计算机,2020,37(12):1-5. [WANG J, WANG J P, SUN W H, et al. A neuromorphic hardware design of a spiking convolutional neural network [J]. Microelectronics & Computer, 2020, 37(12): 1-5.]

用于脉冲卷积神经网络的神经形态处理 VLSI 架构设计

汪 晶,王君鹏,孙文昊,陈 松
(中国科学技术大学,安徽 合肥 230026)

摘 要: 传统的卷积神经网络在训练和识别阶段通常都需要用高能耗的 GPU,无法应用到需要小型低功耗设备的移动应用场景中. 本文设计了一种用于识别手写体的数字脉冲卷积神经网络神经形态硬件 VLSI 架构,根据脉冲神经网络设计的神经形态硬件系统只有在有输入脉冲到来时硬件才会进行相应处理,从而能达到很低的能耗. 在识别 MNIST 数据集时,卷积神经网络识别精度为 99.0%,使用该神经形态硬件的识别精度能达到 98.46%. 相比于相同硬件结构的传统卷积神经网络,平均能耗大大降低.

关键词: 脉冲神经网络;脉冲卷积神经网络;神经形态硬件;手写体数字识别

中图分类号: TN492 **文献标识码:** A **文章编号:** 1000-7180(2020)12-0001-05

A neuromorphic hardware design of a spiking convolutional neural network

WANG Jing, WANG Jun-peng, SUN Wen-hao, CHEN Song

(University of Science and Technology of China, Hefei 230026, China)

Abstract: In recent years, neural network technology has developed rapidly and reached a level comparable to human recognition in the field of image recognition. Traditional convolutional neural networks usually require GPU with high energy consumption in the training and recognition stage, so they cannot be applied to mobile applications requiring small and low-power devices. This paper presents a neuromorphic hardware architecture of spiking convolutional neural networks for recognizing handwriting numbers. Because the neuromorphic only operates when there is a spike input, it can achieve a low power consumption accordingly. When recognizing the MNIST data set, the accuracy of the traditional convolution neural network is 99.0%, whereas the accuracy of the neuromorphic hardware is 98.46%. Therefore, the neuromorphic hardware achieved a comparable recognition accuracy while lowering down the power consumption greatly compared to CNN with a similar hardware architecture.

Key words: spiking neural network; spiking convolutional neural network; neuromorphic hardware; Handwriting digit recognition

1 引言

近年来,人工神经网络(Artificial Neural Network, ANN)已经成为人工智能领域研究的热门方向. 得益于大数据和运算更加高效的硬件设备(例如

GPU)的出现,人工神经网络技术飞速发展,其中卷积神经网络(Convolutional Neural Network, CNN)在图像识别领域更是达到了能和人类识别水平相媲美的程度^[1]. 但是在卷积神经网络中伴随着大量的乘累加运算,硬件能耗非常大,很难将其应用到嵌入

式设备、小型传感器等低功耗设备中, 用与生物神经网络更加接近的脉冲神经网络 (Spiking Neural Network, SNN)^[2] 设计硬件来实现图像识别, 近年来不断受到国内外学者的广泛关注。

使用脉冲神经网络模型设计的神经形态硬件可以用来实现诸如图像识别, 语音识别, 自动驾驶等任务, 并且拥有相比于卷积神经网络应用硬件能耗低得多的优势。

将经过训练后的卷积神经网络转化为脉冲卷积神经网络后通常伴随着识别精度的下降, 近年来, 不少转换算法和技巧被不断提出^[3-5], 其中转换过程导致的识别精度的下降几乎可以忽略不计。因此, 使用神经形态硬件搭建脉冲卷积神经网络来进行图像识别变得越来越普遍^[6]。

近年来, 很多神经形态硬件电路被设计出来, 例如 IBM 公司设计的 TrueNorth^[7]、Intel 公司设计的 Loihi^[8] 以及鲁汶大学研究团队设计的 ODIN^[9] 等。这些神经形态硬件电路可以用于各种图像识别任务中, 且识别过程中的能耗非常低。另外也出现了很多针对脉冲卷积神经网络设计的数字硬件电路^[10], 在文献^[10]中设计的脉冲卷积神经网络数字硬件电路在进行图像识别时同时具有卷积神经网络高准确度的优势和脉冲神经网络低能耗的优势。本设计采用神经形态硬件搭建脉冲神经网络, 在进行图像识别时, 具有较低的能耗和较快的识别速度。

2 脉冲卷积神经网络知识背景

2.1 脉冲神经网络介绍

在传统的神经网络中, 输入通常是高精度的浮点数, 经过逐层的乘累加操作将结果不断向后传递, 最终在输出层得到代表分类结果的概率值。与传统神经网络不同, 脉冲神经网络处理的是按时间和空间分布的脉冲序列, 所以脉冲神经网络又被称为第三代神经网络。

采用 LIF (Leaky Integrated Fire) 神经元模型的脉冲神经网络通常容易进行硬件实现, 其工作原理是: 当有外界脉冲输入时, 输入脉冲经过突触后在突触后端形成与突触权重成比例的电势, 该电势随后累加到神经元膜电压中。当神经元膜电压超过阈值时, 神经元产生输出脉冲并经过其轴突向后传递, 并且膜电压会被相应地重置。LIF 脉冲神经元模型可以表示成下面的公式^[11]:

$$V(t) = V(t-1) + \sum_i w_i s_i(t-1)$$

$$\begin{cases} s(t) = 1, V(t) > V_{th} \\ s(t) = 0, V(t) < V_{th} \end{cases} \quad (1)$$

2.2 脉冲卷积神经网络转换算法

通常, 脉冲神经网络的学习算法分为三类, 分别是有监督学习算法、无监督学习算法以及将经过训练后的卷积神经网络转换为脉冲神经网络。通过适当的转换算法, 可以将训练后的卷积神经网络转换成脉冲卷积神经网络, 精度损失几乎可以忽略不计。

如果将经过训练后的卷积神经网络直接转换成脉冲卷积神经网络, 这样得到的脉冲卷积神经网络虽然在进行图像识别时延时很低, 但是精度同卷积神经网络相比大幅度下降。使用权重归一化转换算法^[5], 能够有效降低上述影响。近期, 普渡大学的研究团队对权重归一化算法进行改进^[4], 并且成功将 VGG^[12], ResNet^[13] 等规模较大的神经网络转换为对应的脉冲卷积神经网络并且精度损失很低, 本设计采用文献^[4]提到的脉冲归一化算法。

本设计使用 Lenet-5 卷积神经网络模型, 通常在训练前需要对该卷积神经网络添加一些约束条件^[5]才能使转化后的脉冲卷积神经网络尽可能接近原始的卷积神经网络: (1) 使用 ReLu 激活函数; (2) 池化层使用平均值池化; (3) 不引入偏置项。

卷积神经网络如果使用 ReLu 激活函数, 得到的激活值非常类似于脉冲神经网络中神经元的平均放电频率; 池化层使用平均值池化以及不引入偏置项的原因是更容易进行硬件实现, 并且采用这些约束对分类精度的影响可以忽略不计^[5]。

3 神经形态硬件整体 VLSI 架构

3.1 神经形态处理核设计

本设计采用的神经形态处理核如图 1 所示, 其原型是来自于数字脉冲神经网络处理器 (Online-learning Digital Spiking Neural Network Processor, ODIN)^[9], 本设计对其进行了一些修改, 删去了与片上学习功能相关的突触权重更新模块, 以应用于脉冲卷积神经网络的设计。

整个神经形态处理核通过输入地址事件表示总线 (Address Event Representation, AER) 和输出 AER 总线与外界进行通信。AER 总线使用四握手异步通信机制, 被广泛使用在神经形态硬件中实现模块之间的脉冲数据传输。如果采用 FPGA 实现该设计, 则可以通过选择配置时钟 IP 核生成时钟信号。控制器对输入信号进行解码, 如果外界输入的是普通的脉冲数据, 将其临时存入调度器的 fifo 中, 等

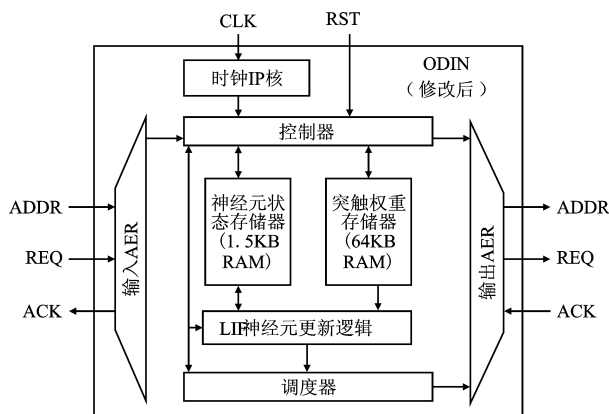


图1 神经形态处理核框图

该硬件模块“空闲”后,即没有外界信号传入该模块,再将 fifo 中的脉冲数据取出进行处理;该硬件模块使用 1.5 KB 的存储器存储 256 个神经元的属性参数,64 KB 的存储器存储所有 256×256 个突触的权重值(每个权重采用 8 bit)。LIF 神经元更新逻辑主要分时复用实现 256 个神经元膜电压的更新,阈值比较和放电,将放电神经元的地址逐一通过输出 AER 模块将脉冲事件输出到外界。

以下简单介绍本设计采取的优化措施及其所带来的影响。为了实现脉冲数据的层间传递,需要在神经形态硬件中加入多个输入 AER 模块和输出 AER 模块,该方法虽然能搭建出脉冲卷积神经网络并且能够进行正常工作,但是由于采用这种方案会产生不同的神经形态处理核,不利于设计的管理。所以,如图 2 所示,在设计中加入发射器将某个神经形态处理核的输出脉冲事件以广播的方式发送到下一层的相应神经形态处理核中去,从而实现一对多的通信;在设计中加入接收器,接收上一层相应的神经形态处理核输出的脉冲事件,通过公平仲裁,将该层的神经形态处理核与上一层的某个神经形态处理核进行连接,从而实现多对一的通信。每个神经形态处理核中包含的 AER 模块由多个减少到一个,减小设计的复杂性,也便于管理以及对该脉冲卷积神经网络的规模和功能进行扩展。

3.2 脉冲卷积神经网络搭建

本设计使用神经形态处理核搭建简化的 Lenet-5 脉冲卷积神经网络,网络的基本结构为:第一层由 6 个卷积核组成;第二层为 2×2 池化;第三层是部分连接层,由 60 个卷积核进行部分连接,最终产生 16 个输出;第四层为 2×2 池化;第五层是全连接卷积层,有 120 个输出,最后一层为全连接层,直接连接到 10 个输出从而得到分类结果。

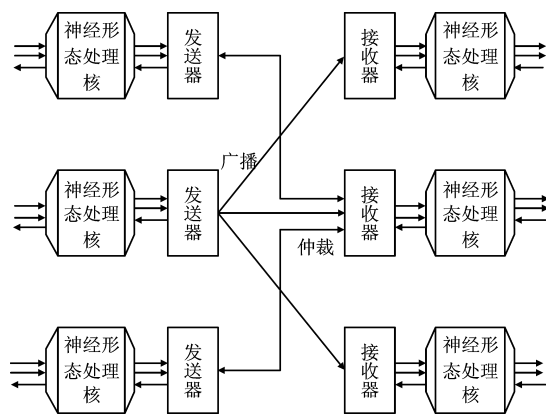


图2 神经形态处理核之间通信硬件逻辑的简化

为了实现相邻层的神经元之间的通信,我们需要加入接收器模块以实现每个神经形态处理核与上一层的多个神经形态处理核之间的通信,同时我们也需要加入发射器模块以实现每个神经形态处理核与下一层的多个神经形态处理核之间的通信。其次,在接收器中我们还需要加入仲裁器对输入其中的多个脉冲输入事件进行仲裁,从而在同一时刻同时有多个脉冲输入事件传入接收器时,接收器只将合适的脉冲输入事件送入到相应的神经形态处理核中。我们使用较普遍的 round-robin 仲裁器^[14],因为我们需要公平的对上一层的每个神经形态处理核输出的脉冲事件进行仲裁以保证通信数据流的吞吐率正常,并且使用该仲裁器能够保证仲裁结果的公平性。

本设计由于是采用图 1 所述的神经形态处理核搭建脉冲卷积神经网络,在完成 Lenet-5 卷积神经网络的训练和到脉冲卷积神经网络的转换以及单个神经形态处理核的设计之后,现在的问题是如何完成硬件模块和卷积神经网络之间的“映射”,即如何组织神经形态处理核来实现原始卷积神经网络中的卷积层、池化层等操作。

本设计采用一个神经形态处理核对应实现原始卷积神经网络的一行卷积操作,这样简化了模块之间的连线逻辑,便于硬件实现。硬件实现和原始卷积神经网络之间的对应关系见图 3。其中 NS_{11}^H 模块对应实现原始卷积神经网络第一层(卷积层)的第一行卷积操作的神经形态处理核, NS_{12}^H 模块对应实现原始卷积神经网络第五层(全连接层)功能的神经形态处理核,其他类似。如图 3 所示, NS_{11}^H 模块实现原始卷积神经网络第一层卷积层的 6 个卷积核与输入图像之间的卷积操作,得到 6 个一行的激活值数据,每个激活值数据对应于 NS_{11}^H 模块中相应神经元的平

均放电频率. 平均值池化层的操作类似于所有权重为 1, 步长为 2 的卷积层卷积操作. 最后一层的全连接操作可以用一个神经形态处理核实现.

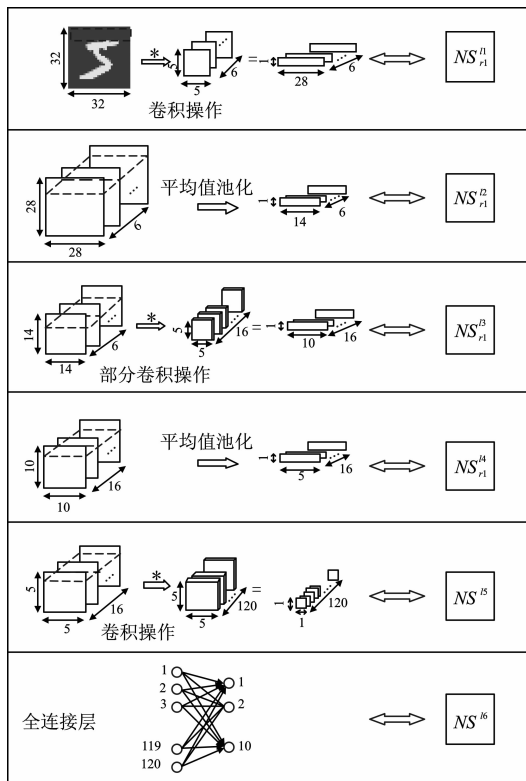


图 3 卷积神经网络与神经形态处理核之间映射关系

按照硬件实现与原始 Lenet-5 卷积神经网络模型之间的对应关系, 我们就可以将神经形态处理核连接起来, 实现最终的脉冲卷积神经网络神经形态硬件, 见图 4.

在进行手写体数字识别时, 经过 Matlab 处理将图片中每个像素转化为平均放电频率与该像素灰度值成比例的脉冲序列, 从而将手写体数字图片转化为对应的脉冲序列二维矩阵, 然后按照时间先后顺序将相应时刻这些脉冲序列之中对应有脉冲的像素地址通过图 4 中的输入 AER 输入到整个脉冲卷积神经网络中. 通过控制器对脉冲地址进行解码, 并根据该脉冲地址与原图片的位置将其送到第一层的相应神经形态处理核中去. 第一层的神经形态处理核对应处理卷积神经网络的第一个卷积层操作, 第二层神经形态处理核对应处理第一个池化层操作, 以此类推. 最后一层的神经形态处理核对应卷积神经网络的最后一层输出层, 所以该模块通过 AER 输出的脉冲事件作为整个识别过程的识别结果.

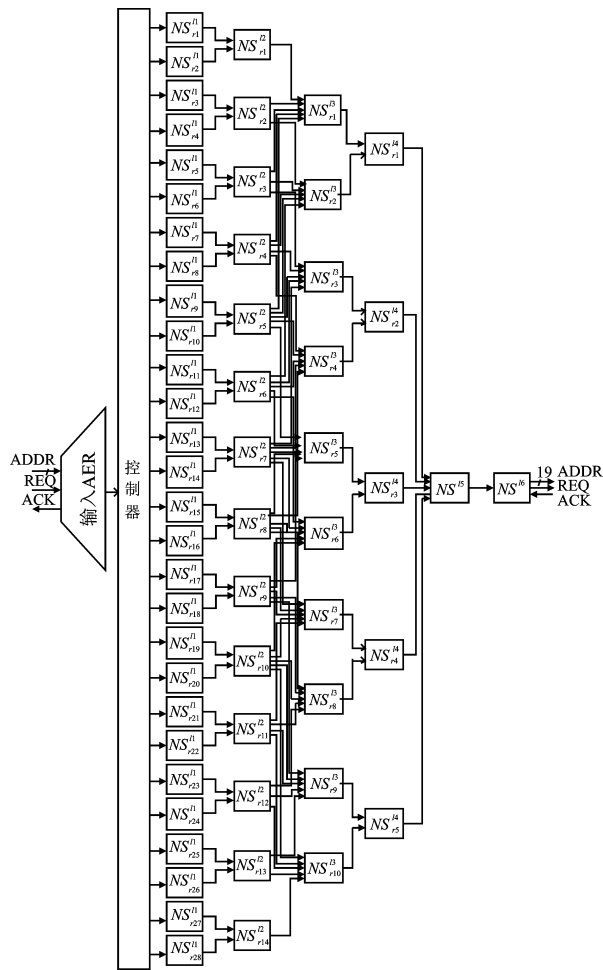


图 4 脉冲卷积神经网络神经形态硬件整体硬件架构

4 电路仿真和测试结果

本文描述的脉冲卷积神经网络神经形态硬件使用 Verilog 硬件描述语言进行设计, 同时基于 SMIC40nm 工艺库, 采用高阈值库来降低静态功耗. 为了能在 DC 综合时得到较为准确的功耗分析结果, 综合前对本设计进行了 VCS 仿真, 对 MNIST 数据集进行识别并得到电路中相关信号的翻转率统计信息. 在时钟频率为 200 MHz 的情况下进行功耗分析, 综合得到的结果与以往设计的脉冲卷积神经网络进行对比^[10].

表 1 是本设计与以往设计的脉冲卷积神经网络以及相同结构的卷积神经网络的电路综合结果和性能的对比. 从图中看出, 相同结构的脉冲神经网络的准确度与卷积神经网络相比都下降了 0.5% 左右, 但功耗与卷积神经网络相比都有了不同程度的降低. 另外, 本文设计的脉冲卷积神经网络的识别速度平均为 83.33 帧每秒, 而在文献[10]中设计的脉冲卷积神经网络的识别速度平均为 12.82 帧每秒.

本设计与以往设计的脉冲卷积神经网络^[10]相比,在面积和准确度都相差不多的情况下,识别速度提高了大约 6.5 倍。

表 1 各硬件电路的综合结果与性能对比

网络	面积 /mm ²	总功耗 /mW	准确度 /%
卷积神经网络 ^[10]	1.42	77.5	99.00
脉冲神经网络 ^[10]	3.31	0.47	98.43
本设计(共计 59 个 神经形态处理核)	3.17	25.0	98.46

5 结束语

本文介绍了脉冲神经网络的工作原理,重点介绍了如何使用神经形态硬件搭建脉冲卷积神经网络的实现.使用该神经形态硬件在进行手写体数字图像识别任务中可以达到很高的识别精度和很低的能耗,并且识别速度与以往的脉冲神经网络相比提高了大约 6.5 倍。

参考文献:

[1] SZE V, CHEN Y H, YANG T J, et al. Efficient processing of deep neural networks: A tutorial and survey [J]. Proceedings of the IEEE, 2017, 105 (12): 2295-2329.

[2] PONULAK F, KASINSKI A. Introduction to spiking neural networks: Information processing, learning and applications [J]. Actaneurobiologiae experimentalis, 2011, 71(4):409-433.

[3] RUECKAUER B, LUNGU I A, HU Y, et al. Conversion of continuous-valued deep networks to efficient event-driven networks for image classification [J]. Frontiers in neuroscience, 2017, 11: 10. 3389/fnins. 2017. 00682.

[4] SENGUPTA A, YE Y, WANG R, et al. Going deeper in spiking neural networks: VGG and residual architectures[J]. Frontiers in neuroscience, 2019, 13: 10. 3389/fnins. 2019. 00095.

[5] DIEHL U, NEIL D, BINAS J, et al. Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing [C]. IEEE International Joint Conference on Neural Networks, 2015: 1-8.

[6] ESSERA S K, MEROLLAA P A, ARTHURA J V,

et al. Convolutional networks for fast energy-efficient neuromorphic computing [J]. PNAS, 2016, 113 (41): 11441-11446.

[7] AKOPYAN F, SAWADA J, CASSIDY A, et al. Truenorth: Design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip [C]. IEEE transactions on computer-aided design of integrated circuits and systems, 2015, 34(10): 1537-1557.

[8] 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.

[9] FRENKEL C, LEFEBVRE M, LEGAT J D, et al. A 0.086-mm² 12.7-pJ/SOP 64k-synapse 256-neuron on-line-learning digital spiking neuromorphic processor in 28-nm CMOS [J]. IEEE transactions on biomedical circuits and systems, 2018, 13(1): 145-158.

[10] XUE T Z, LIU B C, CHEN S. A Hardware Implementation of a Spiking Convolutional Neural Network [J]. Microelectronics & Computer, 2019, 36 (10): 37-41.

[11] ABBOTT L F. Lapicques introduction of the integrate-and-fire model neuron [J]. Brain research bulletin, 1999, 50(5-6): 303-304.

[12] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [C]. ICLR, 2014: 1 - 14.

[13] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770 - 778.

[14] DALLY W J, TOWLES B P. Principles and practices of interconnection networks [M]. Elsevier, 2004: 354-357.

作者简介:

汪 晶 男, (1995-), 硕士研究生. 研究方向为脉冲神经网络神经形态硬件设计.
E-mail: wj50678@mail.ustc.edu.cn

王君鹏 男, (1997-), 硕士研究生. 研究方向为基于忆阻器的神经形态系统可靠性设计.

孙文昊 男, (1997-), 硕士研究生. 研究方向为存算一体架构设计.

陈 松 男, (1979-), 博士, 副教授. 研究方向为脑启发的新型计算架构设计及其自动化, VLSI 计算机辅助设计.