文章编号：2095-4980(2025)02-0170-05

基于“存算一体”的卷积神经网络加速器

卢莹莹[a,b,c]，孙翔宇*[a,b]，计炜梁[a,b]，邢占强[a,b]

(中国工程物理研究院 a.电子工程研究所，四川 绵阳 621999；b.微系统与太赫兹研究中心，四川 成都 610200；
c.研究生院，北京 100088)

摘　要：基于冯·诺伊曼架构的卷积神经网络(CNN)实现方案难以满足高性能与低功耗的要求，本文设计了一种基于“存算一体”架构的卷积神经网络加速器。利用可变电阻式存储器(RRAM)阵列实现“存算一体”架构，采用高效的数据输入管道及硬件处理单元进行大批量图像数据的处理，实现了高性能的数字图像识别。仿真结果表明，该卷积神经网络加速器有着更快的计算能力，其时钟频率可达100 MHz；此外，该结构综合得到的面积为300 742 μm^2，为常规设计方法的56.6%。本文设计的加速模块在很大程度上提高了CNN加速器的速率并降低了能耗，仿真结果对高性能神经网络加速器设计有指导意义和参考作用。

关键词：存算一体；卷积神经网络(CNN)；加速器；输入管道；处理单元

中图分类号：TN79　　**文献标志码：**A　　**doi：**10.11805/TKYDA2023242

Convolutional Neural Network accelerator based on computing in memory

LU Yingying[a,b,c]，SUN Xiangyu*[a,b]，JI Weiliang[a,b]，XING Zhanqiang[a,b]

(a.Institute of Electronic Engineering，China Academy of Engineering Physics，Mianyang Sichuan 621999，China，
b.Microsystem & Terahertz Research Center，China Academy of Engineering Physics，Chengdu Sichuan 610200，China，
c.Graduate School of China Academy of Engineering Physics，Beijing 100088，China)

Abstract：The implementation scheme of Convolutional Neural Network(CNN) based on Von Neumann architecture is difficult to meet the requirements of high performance and low power consumption. Therefore, a CNN accelerator based on storage-computing integrated architecture is designed. By using the circuit structure of Resistive Random Access Memory(RRAM) to realize the storage-computing integrated architecture, and using efficient data input pipeline and CNN processing unit to process large-scale image data, high-performance digital image recognition is realized. The simulation results show that the CNN accelerator has faster computing capability and its clock frequency can reach 100 MHz; in addition, the area of the structure is 300 742 μm^2, which is 56.6% of that of the conventional design method. The acceleration module designed in this paper greatly improves the speed and decreases the energy consumption of CNN accelerator. It shows guiding significance for the design of high performance neural network accelerator.

Keywords：computing in memory；Convolutional Neural Network(CNN)；accelerator；input pipeline；processing unit

“存算一体”最早可追溯至20世纪60年代，后续的相关研究主要围绕芯片电路设计、系统应用优化等层面展开[1]。近年来，随着人工智能的兴起，“存算一体”得到国内外的广泛研究与应用，适用于可穿戴设备、移动设备、智能家居等应用场景[2-3]。

经典冯·诺伊曼计算架构下，数据存储与处理操作分离，存储器与处理器之间通过总线进行数据传输[4-5]。由于处理器和存储器的内部结构、工艺和封装不同，二者的性能存在很大差别。存储器的访问速度远远跟不上中央处理器的处理速度，形成存储器和处理器之间的“存储墙”，严重制约了芯片的性能提升。此外，数据在存储器与处理器之间的频繁迁移带来严重的功耗问题，称为“功耗墙”[6]。由于“存储墙”和“功耗墙”的存在，

收稿日期：2023-09-02；**修回日期：**2023-10-20

***通信作者：**孙翔宇　email:71573841@qq.com

冯·诺依曼架构很难适用于需要大批量数据处理的人工智能应用场景，而“存算一体”架构能够解决这一难题。

“存算一体”架构将数据存储与计算融合在同一芯片中，极大提高了计算速率和能效，特别适用于人工智能领域。本文采用RRAM实现了“存算一体”架构，打破了传统架构中的“存储墙”和“功耗墙”。此外，本文采用高效的数据输入管道，最大程度上复用输入数据，节省存储资源；同时利用硬件加速单元，使数据处理速度翻倍，实现了高性能的图像识别。

1　卷积神经网络设计

1.1 卷积神经网络基本原理

卷积神经网络是一种广泛用于图像和语音识别等领域的深度学习模型，通过模拟人类视觉系统的工作方式，有效提取图像或语音数据中的特征[7]。卷积神经网络由多个层组成，每个层都有特定的功能和原理[8]：

1) 卷积层：卷积层利用卷积操作对输入数据进行扫描并生成特征图，这些特征图可以捕捉到不同位置上的局部特征。

2) 激活层：激活层对卷积层的输出进行非线性变换，引入非线性因素。常见的激活函数包括ReLU、Sigmoid和Tanh等[9]。

3) 池化层：池化层用于减小特征图的空间维度，降低计算复杂度，并增强特征的鲁棒性。常见的池化操作包括最大池化和平均池化。

4) 全连接层：全连接层是将前面各层的输出进行扁平化，通过学习权重和偏差进行分类或预测[10−11]。

总之，卷积神经网络通过卷积、激活、池化和全连接等层的组合，能够从原始数据中自动学习并提取出抽象的特征表示。这些特征表示在图像分类、目标检测、语音识别等任务中发挥重要作用[12−13]。

1.2 卷积神经网络模型选择

本文数据集为MNIST手写数字数据集，共有70 000张图像，其中训练集60 000张，测试集10 000张。所有图像都为28×28的灰度图像，每个数据点取值为0到1范围内的8位浮点数，用8位二进制存储。

采用卷积神经网络模型进行手写数字识别，模型参数如表1所示。

表1 CNN结构参数

Table1 Structural parameters of CNN

network	input	kernel_Num	kernel_size	output
Conv1	28×28	3	5×5	3×24×24
pool1	3×24×24	–	2×2	3×12×12
Conv2	3×12×12	2	5×5	2×8×8
pool2	2×8×8	–	2×2	2×4×4
fully connected	2×4×4	–	–	1×10

表1中network列表明该CNN网络由第一层卷积、第一层池化、第二层卷积、第二层池化、全连接层组成。本文设计中，卷积层和全连接层由“存算一体”芯片实现，池化层由先入先出队列(First In First Out，FIFO)实现。基于RRAM的5层卷积神经网络硬件实现如图1所示。

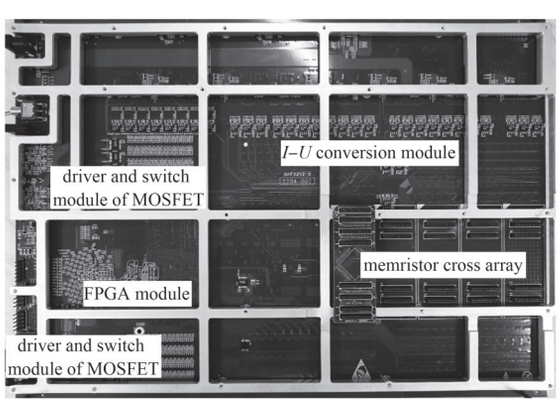

Fig.1 Intelligent computing hardware platform of five−layer convolution neural network based on RRAM

图1 基于RRAM的五层卷积神经网络智能运算硬件平台

2　“存算一体”架构设计

“存算一体”的核心思想是在存储单元内部进行算法嵌入，使数据流动的过程成为输入数据和权重在模拟域做点乘运算的过程，即卷积运算[14−15]。由于卷积运算是深度学习算法的核心组成单元，因此“存算一体”非常适合深度学习[16]。

如图2所示，卷积神经网络模型的权重可映射为存储阵列的电导值，而图像数据则作为电压并行加载进入阵列内部，然后以模拟方式进行输入电压和权重电导的乘法操作，最后进行电流求和生成输出向量[17−18]。

根据RRAM器件结构，可得到卷积运算结果为：

$$\begin{bmatrix}U_1 U_2 U_3\end{bmatrix}\begin{bmatrix}G_{11} & G_{12} & G_{13}\\ G_{21} & G_{22} & G_{23}\\ G_{31} & G_{32} & G_{33}\end{bmatrix}=\begin{bmatrix}I_1 & I_2 & I_3\end{bmatrix} \tag{1}$$

权重映射为电导值，输入数据则映射为电压，数据流动的过程就是点乘运算的过程。具有以下优势：

1) 减少不必要的数据搬运，可降低功耗；

2) 存储单元直接参与计算，可提高性能；

3) 节约大量D触发器占用的芯片面积。

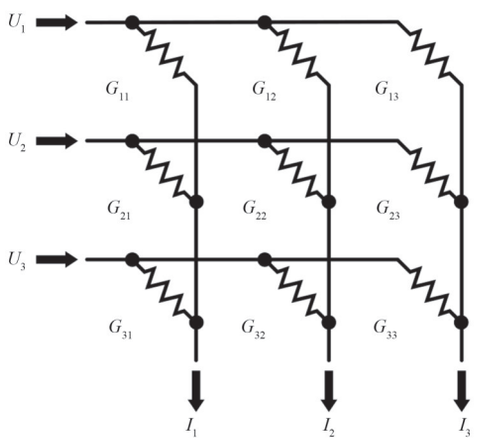

Fig.2 Schematic diagram of RRAM

图2 RRAM存储阵列示意图

3 性能优化方法

性能优化围绕数据复用和流水线设计展开，其中，对于数据复用，本文主要针对图像数据从片外存储器读至片内SRAM这一过程进行优化设计；对于流水线设计，采用并行转串行模块、忆阻器阵列和移位相加模块构成三级流水线优化时序。

3.1 高效数据输入管道

高效的数据输入管道对于提高系统性能非常重要。通过参数化设计，可根据不同的场景配置像素矩阵大小和卷积核大小，使该模块具备通用性，可在顶层改变参数适应不同的应用场景。

通过控制静态随机存取存储器(Static Random-Access Memory，SRAM)数据流，可使该模块在读取数据的同时进行SRAM写操作，将下一行图像数据存储到SRAM内部，如图3所示。这种方法能够大量复用原始图像数据，不仅提高了CNN网络输入数据速率，还节约了大量的存储资源。

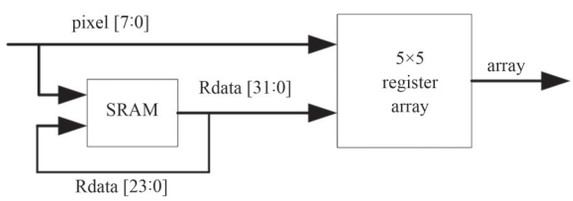

Fig. 3 Schematic diagram of data input pipeline

图3 数据输入管道示意图

3.2 CNN硬件处理单元

在卷积层和全连接层中，采用“存算一体”架构与流水线相结合的方式进行乘累加运算。

如图4所示，移位寄存器将数据转换为串行的8位数据，依次送入RRAM。在卷积层中，每次滑窗取5×5的像素，RRAM同时进行25行乘法操作，并依次输出8个时序的乘累加结果。为获得最终的输出结果，采用移位加法器对忆阻器芯片的输出数据进行移位相加操作。

为进一步提高计算速率，设计中采用了流水线结构。即在进行乘累加运算的同时，可通过移位加法器进行移位相加操作。这种流水线结构不仅提高了电路吞吐量，还提高了系统允许的时钟频率。

在池化层采用平均池化的方式。池化层结构见图5。池化操作需要判断数据的行列位置，如果是偶数行，则将一行数据缓存到FIFO队列中；如果是奇数行，则从FIFO中读取缓存的偶数行数据，并在该行的奇数列进行计算。计算结果是窗口内数据之和经过截位操作，即除以4，实现平均池化。该模块将缓存和计算分离，可提高系统的计算效率。

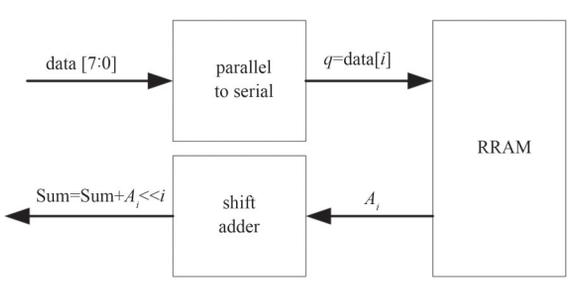

Fig.4 Schematic diagram of three-stage pipeline structure

图4 三级流水线结构示意图

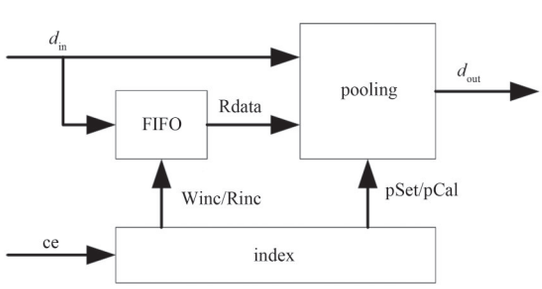

Fig.5 Structure of pooling module

图5 池化模块结构

4 CNN芯片仿真分析

4.1 处理单元准确性分析

对手写数字1~9进行图像识别，全连接层输出结果如图6所示。以数字“2”为例，根据图6可知，数字“2”

的Result2为最大值，代表经过卷积神经网络计算得到的识别结果是数字“2”，与输入图像结果一致，说明CNN芯片计算功能正确。

number	0	1	2	3	4	5	6	7	8	9
result 0	7	0	2	−2	−3	−2	1	−2	−2	−4
result 1	−8	1	3	−9	−4	−23	−6	−10	−13	−4
result 2	0	0	10	−2	−3	−3	0	4	6	−2
result 3	0	0	4	2	0	−9	−4	6	6	1
result 4	−4	0	−7	0	7	−12	−5	−16	−20	3
result 5	2	0	−4	0	0	10	4	−1	0	0
result 6	1	0	−2	−3	−2	6	7	−20	−22	−6
result 7	−3	−1	−12	−12	−1	−6	−7	14	6	0
result 8	0	0	2	0	−3	1	1	5	16	0
result 9	0	0	−14	−6	4	0	−5	3	3	5

Fig.6 Calculation results of fully connection layer

图6 全连接层计算结果

4.2 处理单元速率分析

对比常规的“存算一体”架构[19−20]，本文加入了数据复用和流水线设计等相关优化措施。在进行速率分析和面积分析时，采用相同工艺库及约束条件进行两种实现方式的逻辑综合。

表2为优化前(time 1)和优化后(time 2)的时序参数。常规设计中，数据搬运时间较长，且各级计算单元需等待上一级处理完成之后再进行下一步操作，系统性能受到限制。在满足时序收敛的前提下(slack大于0)，常规设计方法的时钟周期为20 ns，最高工作频率为50 MHz。优化设计后，高效的数据输入管道和流水线设计显著减少了数据搬运和计算时间，系统时钟频率提高至100 MHz。

表2 优化前后时序参数

Table2 Time series parameters before and after optimization

parameter	time 1/ns	time 2/ns
clock CK (rise edge)	20.00	10.00
clock network delay (ideal)	20.00	10.00
EDFFTRXL	20.00	10.00
library setup time	−0.78	−0.47
data required time	19.22	9.53
data arrival time	−16.59	−8.21
slack	2.63	1.32

4.3 处理单元面积分析

电路逻辑的改变会导致组合逻辑单元和时序逻辑单元的数量发生变化。表3为优化前(area 1)和优化后(area 2)芯片逻辑单元的数量和面积参数。

通过对比可知，优化设计中采用了高效输入管道减小存储单元的面积，并引入“存算一体”芯片减小了计算单元的面积。经过设计优化后，逻辑单元的总面积约为300 742 μm^2，仅为优化前面积的56.6%。此外，在相同的芯片制程下，更少的电子元件和电路连接意味着更少的开关动作和信号传输，从而降低了动态功耗。同时，“存算一体”芯片本身也具有较低的功耗，因此优化后一定程度上降低了加速模块功耗。

表3 优化前后芯片逻辑单元数目

Table3 Number of chip logic units before and after optimization

parameter	area 1/μm^2	area 2/μm^2
number of combinational cells	6 939	6 160
number of sequential cells	721	577
net interconnect area	83 286	64 917
total cell area	448 155	235 825
total area	531 441	300 742

综上所述，优化后的卷积神经网络加速器在保持功能正常的同时，实现了更高的性能和更小的物理面积。即优化后的设计能够以更快速率、更低功耗完成目标任务。

5 结论

本文基于RRAM“存算一体”架构，通过采用高效的数据输入管道及硬件处理单元，实现了高性能的数字图像识别。根据实验数据，该加速器工作频率成功提升了一倍，且面积缩减为优化前的56.6%，表明本文设计的基于“存算一体”架构的卷积神经网络加速器在性能和功耗方面具有显著优势；同时也表明本文的研究成果对于高性能神经网络加速器的设计具有指导意义和参考价值。

参考文献：

[1] 张一迪. 阿里达摩院存算一体AI芯片打破存储与计算的"隔阂"[N]. 中国电子报, 2021−12−10(007). (ZHANG Yidi. Ali Dharma Institute's integrated AI chip breaks the "barrier" between storage and calculation[N]. China Electronic News, 2021−12−10.)

[2] 李锟,曹荣荣,孙毅,等. 基于忆阻器的感存算一体技术研究进展[J]. 微纳电子与智能制造, 2019,1(4):87−102. (LI Kun,CAO Rongrong,SUN Yi,et al. Research progress on the fused technology of sensing, storage and computing based on memristor[J]. Micro/Nano Electronics and Intelligent Manufacturing, 2019,1(4):87−102.) doi:10.19816/j.cnki.10−1594/tn.000033.

[3] 李雅琪,温晓君. 存算一体化的发展现状挑战与对策建议[J]. 互联网经济, 2020(4):15−17. (LI Yaqi,WEN Xiaojun. Development status,challenges and countermeasures of deposit and calculation integration[J]. Digital Economy, 2020(4):15−17.)

[4] 高玉光. 基于RRAM的存算一体芯片外围电路和并行计算设计[D]. 西安:西安电子科技大学, 2022. (GAO Yuguang.

Peripheral circuit and parallel computing design of memory and computing integrated chip based on RRAM[D]. Xi'an,China: Xidian University, 2022.)

[5] SEBASTIAN A,LE GALLO M,KHADDAM-ALJAMEH R,et al. Memory devices and applications for in-memory computing[J]. Nature Nanotechnology, 2020,15(7):529-544. doi:10.1038/s41565-020-0655-z.

[6] HUANG Xiaohe,LIU Chunsen,JIANG Yugang,et al. In-memory computing to break the memory wall[J]. Chinese Physics B, 2020,29(7):078504. doi:10.1088/1674-1056/ab90e7.

[7] 张松兰. 基于卷积神经网络的图像识别综述[J]. 西安航空学院学报, 2023,41(1):74-81. (ZHANG Songlan. A review of image recognition based on convolutional neural network[J]. Journal of Xi'an Aeronautical University, 2023,41(1):74-81.)

[8] 陈群贤. TensorFlow下基于CNN卷积神经网络的手写数字识别研究[J]. 信息记录材料, 2022,23(9):159-161. (CHEN Qunxian. Research on handwritten numeral recognition based on CNN convolutional neural network under TensorFlow[J]. Information Recording Materials, 2022,23(9):159-161.) doi:10.16009/j.cnki.cn13-1295/tq.2022.09.056.

[9] 卢金波. 基于卷积神经网络的图像检测方法研究[D]. 西安:西安理工大学, 2022. (LU Jinbo. Research on image detection method based on convolutional neural network[D]. Xi'an,China:Xi'an University of Technology, 2022.)

[10] 林朋雨,郭杰. 基于FPGA的卷积神经网络加速优化方法[J]. 计算机仿真, 2022,39(7):371-374,450. (LIN Pengyu,GUO Jie. FPGA-based accelerated optimization method of convolutional neural network[J]. Computer Simulation, 2022,39(7):371-374, 450.) doi:10.3969/j.issn.1006-9348.2022.07.071.

[11] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017,60(6):84-90. doi:10.1145/3065386.

[12] SIMONYAN K,ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[DB/OL]. arXiv, 2014-11-18.

[13] 周伟. 基于卷积神经网络的多聚焦图像融合算法研究[D]. 南京:南京邮电大学, 2022. (ZHOU Wei. Research on multi-focus image fusion algorithm based on convolutional neural network[D]. Nanjing, China: Nanjing University of Posts and Telecommunications, 2022.)

[14] 陆春帆,刘爽,周洲. 一种新型存算一体电路设计与研究[J]. 科学技术创新, 2022(36):91-94. (LU Chunfan,LIU Shuang,ZHOU Zhou. Design and research of a new type of integrated storage and computing circuit[J]. Scientific and Technological Innovation, 2022(36):91-94.) doi:10.3969/j.issn.1673-1328.2022.36.024.

[15] WANG Yimin,ZOU Zhuo,ZHENG Lirong. Design framework for SRAM-based computing-in-memory edge CNN accelerators[C]// 2021 IEEE International Symposium on Circuits and Systems(ISCAS). Daegu,Korea:IEEE, 2021:1-5.

[16] TAN Fei,WANG Yiming,YANG Yiming,et al. A ReRAM-based computing-in-memory convolutional-macro with customized 2T2R bit-cell for AIoT chip IP applications[J]. IEEE Transactions on Circuits and Systems II:Express Briefs, 2020,67(9):1534-1538. doi:10.1109/TCSII.2020.3013336.

[17] YANG Jiancheng,CUI Aoxin,JIA Song,et al. A configurable computing-in-memory structure based on convolutional neural network[C]// 2021 China Semiconductor Technology International Conference(CSTIC). Shanghai,China:IEEE, 2021:1-3.

[18] LIU Bicheng,GU Shouzhen,CHEN Mingsong,et al. An efficient racetrack memory-based processing-in-memory architecture for convolutional neural networks[C]// 2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications(ISPA/IUCC). Guangzhou, China:IEEE, 2017:383-390. doi:10.1109/ISPA/IUCC.2017.00061.

[19] XUE C X,CHEN W H,LIU J S,et al. Embedded 1 Mb ReRAM-based computing-in-memory macro with multibit input and weight for CNN-based AI edge processors[J]. IEEE Journal of Solid-State Circuits, 2020,55(1):203-215.

[20] MOCHIDA R,KOUNO K,HAYATA Y,et al. A 4M synapses integrated analog ReRAM based 66.5 TOPS/W neural-network processor with cell current controlled writing and flexible network architecture[C]// 2018 IEEE Symposium on VLSI Technology. Honolulu,HI,USA:IEEE, 2018:175-176. doi:10.1109/VLSIT.2018.8510676.

作者简介：

卢莹莹(1999-)，女，在读硕士研究生，主要研究方向为基于存算一体架构的智能芯片设计. email: 291945996@qq.com.

孙翔宇(1988-)，男，博士，副研究员，主要研究方向为微系统集成、MEMS传感器与执行器.

计炜梁(1992-)，男，硕士，助理研究员，主要研究方向为压电MEMS传感器、超声探测算法.

邢占强(1991-)，男，硕士，助理研究员，主要研究方向为压电MEMS传感器、电路设计.