基于忆阻器的LSTM神经网络存内计算架构设计

叶沁桐
（福州大学 物理与信息工程学院，福建 350001）

摘要：阐述一种基于忆阻器阵列的存内计算宏电路，用于加速循环神经网络（RNN）中的长短期记忆神经网络（LSTM）。长短期记忆神经网络能够更容易地捕捉深层连接，并改善梯度消失和梯度爆炸现象。传统的LSTM硬件加速器使用乘法器与加法器实现网络中最频繁的点积操作，硬件系统中的计算资源决定了整个网络的加速效率，同时，权重等参数在内存与计算单元之间的通信也带来了较大的功耗。

关键词：集成电路设计，存内计算，忆阻器，神经网络，硬件加速。

中图分类号：TN402，TN60，TP183　文章编号：1674-2583(2023)10-0018-04
DOI：10.19339/j.issn.1674-2583.2023.10.007
文献引用格式：叶沁桐.基于忆阻器的LSTM神经网络存内计算架构设计[J].集成电路应用,2023, 40(10): 18-21.

Design of LSTM Neural Network Hardware Accelerator Based on Memristor

YE Qintong
(School of Physics and Information Engineering, Fuzhou University, Fujian 350001, China.)

Abstract — This paper describes ta memory-resistor array-based in-memory computing macro circuit for accelerating long short-term memory (LSTM) in recurrent neural networks (RNNs). LSTMs, which are recursive neural networks with long short-term memory, are able to more easily capture deep connections and improve gradient vanishing and exploding phenomena. However, traditional LSTM hardware accelerators use multipliers and adders to implement the most frequent dot product operations in the network, which may lead to problems such as overly complex models and overly large parameters, resulting in computational bottlenecks

Index Terms — integrated circuit design, in memory computing, memristors, neural networks, hardware acceleration.

0 引言

近年来，神经网络的发展迅速，在图像识别、语音识别、自然语言处理等领域广泛应用。其中，卷积神经网络常用于计算机视觉领域，而循环神经网络主要应用于语音识别和自然语言处理[1]。随着大数据时代的到来，神经网络被应用于更复杂的情境，对于网络精度的要求越来越高。因此，网络结构变得更加复杂，包含的参数量也增加。因此，神经网络需要更高算力的计算设备。在实际应用中，大多数神经网络由中央处理器（CPU）和图像处理器（GPU）实现。但是，CPU的并行度较低，导致网络训练过程面临较高延迟的问题。尽管GPU具有足够的并行度，但也面临成本高、功耗大的问题[2]。除此之外，有人设计专用集成电路（ASIC）方法来实现神经网络的硬件部署，然而ASIC设计周期长，成本高，并且其电路结构无法随着快速变化的网络结构而变化。现场可编程门阵列（FPGA）由于它高度可配置性，并行性的优点，是一种被广泛应用于在边缘设备实现神经网络推理过程的器件[3]。

然而，上述的几种硬件平台都面临着一个同样的问题：庞大的数据交换带来的巨大的能量消耗。在传统“冯·诺依曼”架构的设备上，内存与处理器是分开的，这使得神经网络的大量数据不得不频繁地在内存与处理器之间传输，这就是著名的“存储墙”问题[4]。因此，本文提出了一种非易失性的存储阵列-忆阻器阵列，在存储神经网络的权重参数的同时，完成权重与输入的点积操作。

1 循环神经网络

1.1 循环神经网络

人工神经网络（ANNs），特别是深度神经网络（DNNs）在图像分类、视频识别、语音识别和自然语言处理等应用中取得了很大的成功。卷积神经网络（CNN）在图像分类任务中被广泛使用。然而，这种前馈神经网络并不适合处理和时间相关的数据序列，如语音、文字和影像。另一类人工神经网络-循环神经网络（RNN），被广泛应用于捕

作者简介：叶沁桐，福州大学物理与信息工程学院；研究方向：集成电路设计。
收稿日期：2022-12-31；修回日期：2023-09-23。

捉序列数据之间的时间依赖性[5]。它能够处理序列数据，并利用之前的计算结果作为本次计算的输入。

1.2 LSTM

然而，梯度消失和梯度爆炸是影响神经网络性能的普遍问题，普通的循环神经网络无法记住超过一定时间的信息。为了解决这个问题，人们提出了长短期记忆神经网络(LSTM)，它已成为最广泛使用的循环神经网络结构。

本文所述的LSTM结构是较为简单的结构[6]，包含两层LSTM网络。与其他版本的LSTM相比，该结构的预测精度在可接受的范围之内。图1展示了LSTM的架构。Embedding层将输入的字符串或文本信息转换为独热码表示的向量x(t)并输出给两层LSTM。每层LSTM中，数据之间进行点积运算。tanh和 代表激活函数。f(t)、i(t)、c(t)和o(t)分别表示遗忘门、输入门、状态门和输出门。通过控制门单元，LSTM能够记住或忘记先验信息c(t-1)、输入x(t)和上一时刻的输出h(t-1)得到一层的输出h(t)。以遗忘门为例，来自上一时刻的输出和当前输入同时传入到 函数中，输出一个在0到1之间的值，输出值越接近0，意味着该信息应该丢弃，而越接近1则表示该数据应该被保留。W和b分别表示权重矩阵和偏置向量。每个门有自己对应的权值矩阵和偏置向量。输入向量x的大小为X，输出向量h、c的大小为H。矩阵W的大小为H × X。在第二层LSTM输出结果h(t)后，通过一个全连接输出层输出预测结果y(t)

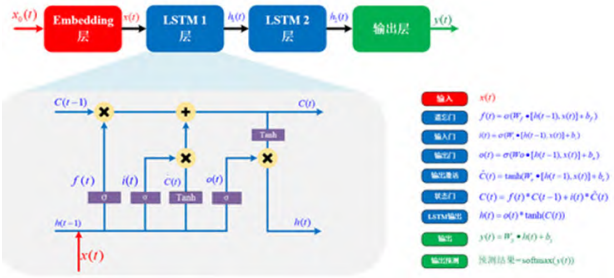


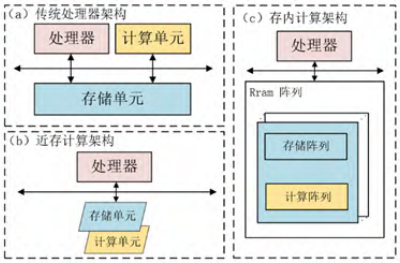
图1 LSTM神经网络结构

2 存内计算与忆阻器件

2.1 存内计算

在传统的计算机系统中，处理单元（如 CPU 和 GPU）和数据存储器（如内存、闪存和磁盘）是分开的。过去的十年中，由于需要处理的数据量的快速增长，处理单元与内存之间的数据交换成为影响计算机系统性能和功耗的关键因素。在这种情况下，内存处理（processing-in-memory，PIM）技术为解决这一问题提供了新的思路。图2展示了目前的几种神经网络处理架构。图2（a）为传统的“冯诺依曼”处理器架构，虽然随着技术的发展，处理器与存储器的性能不断提高。但负责通信的总线传输速度跟不上两者的速度。不仅如此，频繁的数据交换所带来的功耗也是在边缘设备中实现神经网络不能接受的一个问题。基于此，有人提出了近存计算这一概念，近存计算的指的是在接近数据所在的位置对数据进行处理。如图2（b）所示，这种方法以数据为中心，将靠近数据的计算单元耦合在一起，减少数据移动所带来的功耗损失。三维堆叠技术是近存计算架构的真正推动者，它允许通过硅通孔（TSV）将逻辑单元和内存堆叠在一起，这有助于减少内存访问延迟、功耗并提供更高的带宽[7]。

近存计算方法虽然有效减少了数据传输带来的功耗，但本质上还是没有走出传统的冯诺依曼架构。存内计算结构的出现打破了这一现状。如图2（c）所示，它将存储器件直接用于计算操作，这通常意味着需要改变存储阵列，使其支持计算。目前，存内计算的研究主要集中在SRAM、DRAM和其他新型存储器件（如RRAM、PCRAM、MRAM）上。一般来说，基于模拟计算的存内计算设计功耗更低。



（a）传统处理器架构（b）近存计算架构（c）存内计算架构

图2 神经网络处理架构

2.2 忆阻器器件

非易失性的存储器件（Resistive Random Access Memory，RRAM）在断电后，存储的数据也不会丢失。RRAM使用可变的电阻作为基本存储单元来存储信息。其原理图如图3所示，它的结构非常简单，由两层金属电极和一层金属氧化物层组成，其中金属氧化物层夹在金属上电极和金属下电极之间[8]。对于一个RRAM存储单元，用低电阻状态(LRS)和高电阻状态(HRS)来表示逻辑“1”和“0”。通过施加特定极性、幅值和持续时间的外部电压，可以切换RRAM单元的高低阻态。

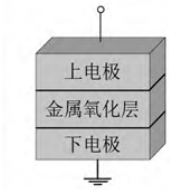


图3 忆阻器单元

2.3 忆阻器阵列架构

根据RRAM单元的功耗、延迟、成本不同，可以将RRAM阵列设计为具有1T1R单元的网格结构，也可以设计为密集的交叉阵列（crossbar）结构[9]。

图4（a）是一种传统的RRAM存储单元结构，其中每个单元都由一个MOSFET晶体管与一个忆阻单元组成。与DRAM类似[10]，激活忆阻器阵列的一行时，只有被选中的1T1R单元会被访问，阵列中的其他单元不受干扰。然而，与DRAM不同的是，RRAM存储器通常在更高的电流下工作，每个存储

单元中的晶体管尺寸更大，从而导致了整个RRAM存储器面积与成本的增加。然而，与其他替代方案相比，“1T1R”设计更节能，读写时间更快。本文提出的忆阻器外围电路是基于1T1R结构的忆阻器阵列来设计的。交叉阵列架构的忆阻器阵列如图4（b）所示，其中所有忆阻单元都是相互连接的，并不存在MOSFET晶体管。RRAM单元被直接夹在顶部和底部电极之间。由于每个单元都减少了晶体管，交叉阵列中的存储单元的理论尺寸可以更小，交叉阵列下方的硅面积可用于其他外围电路，如解码器和驱动器，从而最大化阵列的面积效率。相对于1T1R结构，交叉阵列结构的密度更高，能够提供更快的访问速度和更低的功耗。然而，它的缺点是设计和制造更加复杂，成本也更高。

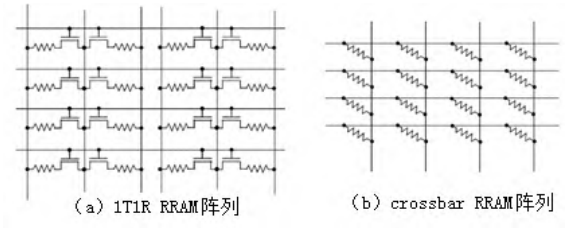


图4 忆阻器阵列架构

3 基于忆阻器件的LSTM存内计算架构

3.1 忆阻器与权重映射

随着RRAM与1T1R架构的出现，RRAM有希望成为一种高效的构建神经网络计算阵列的方法。Presioso等人基于一个12 × 12 RRAM交叉开关忆阻器阵列实现了一个神经网络，成功地将3*3像素的黑白图像分为3类[11]。在构建神经网络时，通常采用一种表示神经元之间相互连接的方法，包括输入层、输出层和多个隐藏层。图5中展示了一个简单的神经网络连接关系，输入层与输出层都包含了两个神经元。输出OUTj可以表示为式（1）。

$$OUT_j = \sigma(\sum_{\forall i} IN_i \times W_{ij}) \quad (1)$$

其中，INi是输入数据，Wij是权重，σ是一个激活函数。可以看出，乘累加操作是神经网络中最常见的一种计算方式。图5还显示了如何使用一个2 × 2的RRAM阵列来实现这种简单的神经网络。当位线（WL）被激活时，MOS开关被打开，通过在字线（BL）上施加一个不引起RRAM阻值变化的模拟电压VBLC来表示输入数据INi；权重Wij被写入交叉阵列中的RRAM单元中，这样流向位线末端的电流I就是矩阵向量乘法的结果。

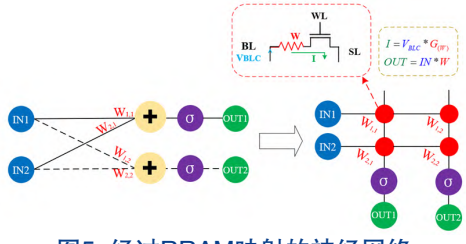


图5 经过RRAM映射的神经网络

3.2 忆阻器外围计算电路设计

基于忆阻器阵列的外围计算电路架构如图6，整个外围计算电路由16个通道组成，每个通道输入4bit信息，经过忆阻器阵列中的4bit权重，16个通道最终得到11bit输出。作为一个4kB的忆阻器阵列的外围电路，电路主要包含了控制模块，电压映射模块，高阻态电流消除模块，电流求和模块，电流量化模块，基准电压产生模块以及数据整合模块。

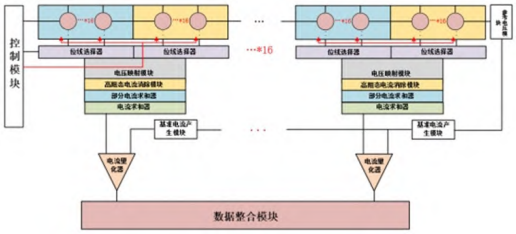


图6 存内计算宏电路架构

各通道的输入均为4bit的数字量，每个4bit的输入([3:0])在控制模块的作用下被分为顺序的2bit（[1:0]和[3:2]）依次输入至电压映射模块中，该模块通过输入的两位数字量（00、01、10、11）产生四种不同的钳位电压输入到忆阻器阵列并作用在相应的位线上，在一个RRAM存储单元中完成一次模拟计算。

在RRAM存储单元中进行一次模拟计算后，代表计算结果的电流IDL会进入高阻态电流消除模块。理想忆阻器的高阻态阻值无限大，高阻态电流近似为零，而实际忆阻器高阻态阻值的大小可能仅为低阻态阻值的几倍，导致高阻态电流不能忽略。因此高阻态电流消除模块被用来消除忆阻器高阻态阻值不够大所导致的较大高阻态电流。两次2bit输入乘4bit权重的乘法计算电流经过部分电流求和模块从二进制补码电流转化为十进制电流，在电流求和模块中相加。至此，一个通道中的4bit输入乘以4bit权重计算完成。乘累加电流最终输入电流量化模块进行量化。根据基准电压模块产生基准电流来将乘累加电流比较量化成对应的数字量。16个通道的计算结果输入到数字整合模块中。得到11bit的乘累加结果，结束一次运算。

3.3 基于忆阻器的LSTM神经网络存内计算架构

图7展示了基于4kB RRAM阵列的存内计算芯片与FPGA构成的LSTM神经网络存内计算架构平台。在FPGA板上模拟了存内计算宏电路的控制电路、LSTM神经网络的Embedding层、用于对存内计算宏电路产生的部分点积批次求和的数字电路、σ激活函数和tanh激活函数。使用FPGA芯片上的片上存储资源作为每一个序列的输入缓冲区。在初始阶段阶段，通过FPGA向RRAM阵列发送内存地址与写权重命令，将权重写入RRAM阵列中。在推理阶段，FPGA向存内计算芯片发送embedding过后的输入数据，并收集输出数据。在收集了一个序列的所有输

出数据后，FPGA执行对应的激活函数。所有序列与层的计算结果被储存在FPGA存储器中，作为下一层的输入。经过多层迭代处理后可以得到最终的输出结果。

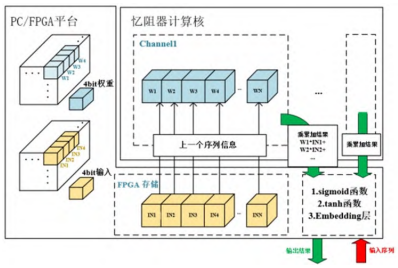


图7 LSTM神经网络存内计算架构

4 仿真结果与分析

表1为本工作相关参数。主要在40nm工艺下实现了一个4kB大小的存内计算宏电路，在4bit权重，4bit输入下可以得到11bit的较高精度输出。该方案实现多位点积操作需要的计算时间T为85ns。

表1 存内计算宏电路参数

	本工作
工艺/nm	40
忆阻器开销/kB	4
输入位数/bit	4
权重位数/bit	4
输出位数/bit	11
T/ns	85ns

为了确认使用于存内计算宏电路的4bit输入与4bit权重的数据精度对LSTM神经网络的影响，基于谷歌语音命令数据集在Tensorflowr框架上进行了软件模型的搭建与训练，训练结果如表2所示。由于将输入与权重量化至4bit时带来的精度差距，LSTM的预测错误率比全精度时增加了5.6%。但同时也减少了模型的计算量与数据量。

表2 LSTM量化错误率

	64bit 输入 64bit 权重	8bit 输入 8bit 权重	4bit 输入 4bit 权重
错误率	10.5%	13.2%	16.1%

5 结语

本文对循环神经网络以及其重要分支LSTM神经网络做了详细介绍，并基于忆阻器阵列设计了一个存内计算宏电路。基于设计的电路提出了一种异构的LSTM神经网络存内计算架构。由于输入数据精度的影响，该系统适合在资源较少，算力较低，对预测结果精度要求不高的应用环境下部署。

参考文献

[1] Tsai H, Ambrogio S, Mackin C, et al. Inference of Long-Short Term Memory networks at software-equivalent accuracy using 2.5M analog Phase Change Memory devices[C]. 2019 Symposium on VLSI Technology. 2019.

[2] Song L, Qian X, Hai L, et al. PipeLayer: A Pipelined ReRAM-Based Accelerator for Deep Learning[C]. 2017 IEEE International Symposium on High Performance Computer Architecture (HPCA). IEEE, 2017.

[3] Cao S, Zhang C, Yao Z, et al. Efficient and Effective Sparse LSTM on FPGA with Bank-Balanced Sparsity[C]. the 2019 ACM/SIGDA International Symposium, ACM, 2019.

[4] Ming, Cheng, Lixue, et al. TIME: A Training-in-Memory Architecture for RRAM-Based Deep Neural Networks[J]. IEEE Transactions on Computer Aided Design of Integrated Circuits & Systems, 2018.

[5] Bank-Tavakoli E, Ghasemzadeh S A, Kamal M, et al. POLAR: A Pipelined/Overlapped FPGA-Based LSTM Accelerator[J]. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2019(99): 1-5.

[6] Greff K, Srivastava R K, J Koutník, et al. LSTM: A Search Space Odyssey[J]. IEEE Transactions on Neural Networks & Learning Systems, 2016, 28(10): 2222-2232.

[7] Gupta S, Imani M, Kaur H, et al. NNPIM: A Processing In-Memory Architecture for Neural Network Acceleration[J]. IEEE Transactions on Computers, 2019.

[8] Wong H S P. Metal－Oxide RRAM[J]. Proceedings of the IEEE, 2012, 100(6): 1951-1970.

[9] Cong X, Niu D, Muralimanohar N, et al. Overcoming the challenges of crossbar resistive memory architectures[C]. IEEE International Symposium on High Performance Computer Architecture, IEEE, 2015.

[10] A. N. Udipi, N. Muralimanohar, N. Chatterjee, R. Balasubramonian, A. Davis, N. P. Jouppi. Rethinking DRAM Design and Organization for Energy-Constrained Multi-Cores[C]. in Proceedings of the International Symposium on Computer Architecture (ISCA), 2010: 75-186.

[11] Prezioso, M, Merrikh-Bayat, et al. Training andoperation of an integrated neuromorphic network based on metal-oxide memristors[J]. Nature, 2015.