

稀疏递归神经网络的可扩展低功耗加速器

金磐石¹, 李俊杰², 王静逸², 李鹏翀³, 邢磊², 李晓栋¹

(1. 中国建设银行股份有限公司, 北京 100034;

2. 建信金融科技有限责任公司, 上海 321004;

3. 浪潮电子信息产业股份有限公司, 山东 济南 250000)

摘要: 利用银行网点内边缘计算设备进行客流分析、安全保护、风险防控等应用日益广泛, 其中 AI 推理芯片的性能和功耗已经成为边缘计算设备选型的一个非常重要的因素。针对递归神经网络由数据依赖性和低数据重用性导致的功耗大、推理性能弱、能效低, 难以在低功耗平台上处理等问题, 利用 FPGA 实现了一种电压可扩展的稀疏循环神经网络 (RNN) 低功率加速器, 并在边缘设计设备上进行了验证。首先, 对稀疏 RNN 进行分析并采用网络压缩的方法设计了处理阵列; 其次, 由于稀疏 RNN 的工作负载不平衡, 引入电压缩放方法以保持低功耗和高吞吐量。试验表明, 该方法可以显著提高系统的 RNN 推理速度并降低芯片的处理功耗。

关键词: RNN; 稀疏; 低功耗; 加速方案

中图分类号: TP393

文献标识码: A

doi: 10.11959/j.issn.2096-8930.2023045

Scalable Low Power Accelerator for Sparse Recurrent Neural Network

JIN Panshi¹, LI Junjie², WANG Jingyi², LI Pengchong³, XING Lei², LI Xiaodong¹

1. China Construction Bank Co., Ltd., Beijing 100034, China

2. Jianxin Financial Technology Co., Ltd., Shanghai 321004, China

3. Inspur Electronic Information Industry Co., Ltd., Jinan, Shandong 250000, China

Abstract: The use of edge computing devices in bank outlets for passenger flow analysis, security protection, risk prevention and control is increasingly widespread, among which the performance and power consumption of AI reasoning chips have become a very important factor in the selection of edge computing devices. Aiming at the problems of recurrent neural network, such as high power consumption, weak reasoning performance and low energy efficiency, which were caused by data dependence and low data reusability, this paper realized a sparse RNN low-power accelerator with scalable voltage by using FPGA, and verifies it on the edge design and calculation equipment. Firstly, the sparse -RNN was analyzed and the processing array was designed by network compression. Secondly, due to the unbalanced workload of sparse RNN, it introduced voltage scaling method to maintain low power consumption and high throughput. Experiments show that this method could significantly improve the RNN reasoning speed of the system and reduce the processing power consumption of the chip.

Keywords: RNN, sparse, low power consumption, acceleration scheme

0 引言

在智慧网点建设中, 通过边缘计算设备可以将智能推理从云端下沉到边缘节点。由终端设备负责感知和采集用户行为及周围环境数据, 边缘节点利用云端下发的模型对采集用户和现场环境数据进行分析, 可快速实现用户身份

识别、行为分析、轨迹跟踪、热点区域分析、流量分析、网点设备智能管控、抵质押物全流程跟踪、金库智慧化管理等创新应用。

在智慧网点的部署过程中, 低功耗边缘设备由于功耗低、噪声小, 对部署场地的要求较低, 因此得到了更为广泛的部署应用, 但是这也限制了这类设备的边缘推理和计

算能力。如何在提高推理能力的同时降低设备的功耗，就成了一个亟须解决的问题。

具体的突破口在于循环神经网络(RNN)的效率提升。RNN 是一种广泛应用的神经网络结构，其主要特点包括神经元之间的内部反馈和前馈突触连接。相对于传统的前馈神经网络，在处理时间相关任务如语音识别、自然语言处理等方面，RNN 具有明显的优势，因为其内部单元能够展现出动态的时间行为。尽管 RNN 在处理序列数据方面具有巨大潜力，但在过去几十年中，对其研究相对有限。这主要是因为 RNN 的计算复杂性较高，因此难以成为边缘计算和端侧机器学习^[1-2]的主流工具。本文对 RNN 的计算过程进行了优化和能效提升。

1 稀疏 RNN 处理阵列的设计

1.1 稀疏 RNN 分析

矩阵向量乘法($M \times V$)是 RNN^[3]中的一个基本构建块。在这些计算过程中，内存访问的数据比以往多，通常 RNN 的矩阵相对较大，这使得内存成为瓶颈。为了降低与大量突触相关的计算复杂度和大存储，业界提出了几种有效的技术，使 RNN 稀疏而不损失网络精度，特别是网络压缩技术。

深度压缩方法作为先进的技术之一，由宋韩等^[4]在 2015 年提出的。在深度压缩中，3 个阶段的管道共同作用，包括修剪、训练量化和霍夫曼编码，以减少网络中的突触权值的数量。首先，通过删除冗余连接来修剪网络，只保留信息最丰富的连接；接下来对权值进行量化，使多个连接共享相同的权值，因此只需要存储有效权值和索引；最后应用霍夫曼编码来利用有效权值的有偏分布。通过上述步骤，最终将矩阵向量乘法转化为稀疏矩阵向量乘法 SPMV。

经过网络压缩后权重矩阵分布稀疏，如果按照传统的矩阵向量乘法方法计算，会导致性能下降^[5]，如图 1 所示。

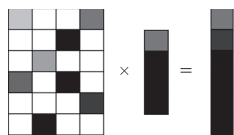


图 1 传统的矩阵向量乘法方法

第一，根据权重矩阵中每一行中的非零元素的位置来计算偏移地址，从而可以加载输入向量 x 中相应的元素，增加了计算时延；第二，对输入向量的不规则访问会导致比率的降低，并增加时延^[6]。

本文提出的体系结构放弃通过输入向量与矩阵的每一行之间的点积运算来获得输出向量的每个元素的思想。相反，矩阵的每一列都由输入向量的相应元素和添加的逐列进行加权，如图 2 所示。

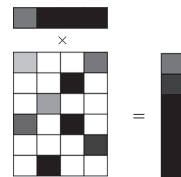


图 2 优化的矩阵向量乘法

1.2 处理稀疏 RNN 数组

由于稀疏 RNN 中最重要的计算是乘法和累加运算，本文提出了一种适用于稀疏矩阵乘法的加速单元。该加速单元的核心单元是处理元件 (PE)，这是一个 16 位的数据处理单元。由于每个处理阵列都包含大量的 PE 单元，所以每个 PE 的结构都应该尽可能简单，如图 3 所示。每个 PE 包括输入寄存器、权重寄存器、输出寄存器和一个临时数据寄存器，以及一个简单地执行一个乘法和累加操作的 ALU。

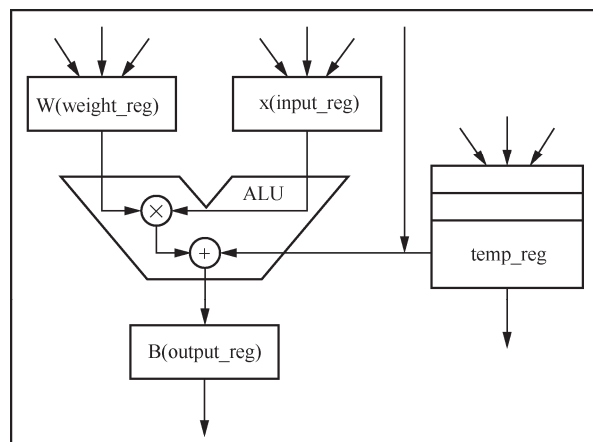


图 3 处理单元的结构

图 4 为处理阵列的架构。RCA 是一种可重构可配置的计算阵列，4 个 RCA 一起协同工作，共享缓冲区 Bank1、Bank2。RCA 的网络权重由 Weight SRAM 处理，进行压缩是效率提升的关键因素之一。RCA 的核心单元是 PEA 处理阵列。

处理阵列包括一个二维处理元素 (PE) 阵列。此外，通过评估算法映射效率和芯片面积开销之间的权衡，PE 阵列的大小可以灵活地放大或缩小。本文在每个 RCA 中采用 6×8 的 PE。

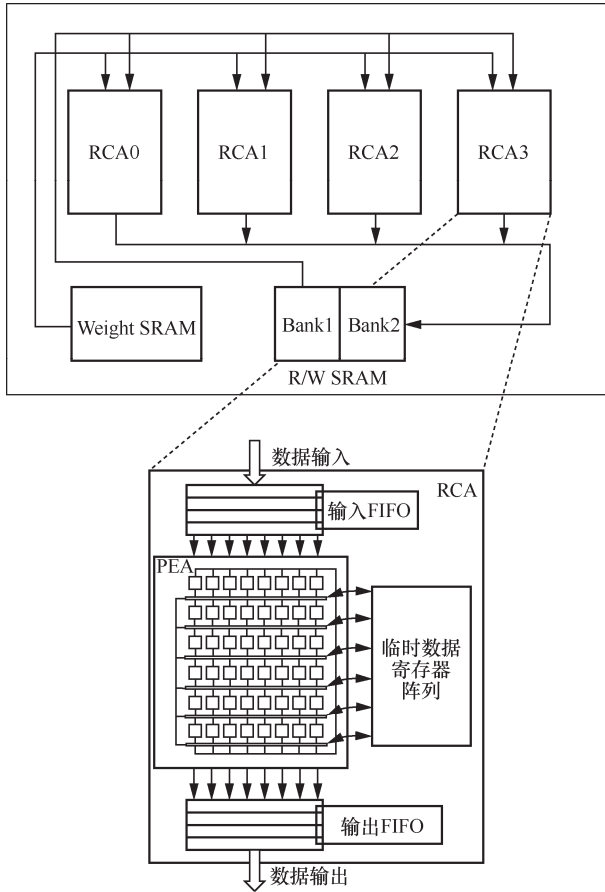


图4 处理阵列的架构

工作负载感知单元收集执行复杂度信息,即每个稀疏权值矩阵的非零元素的数量。

电压-频率(VF)缩放控制器根据执行复杂度来描述所需的执行时间和计算频率。

缓冲器用来存储输入和输出数据(与RCA紧密耦合)。由于深层RNN中每一层的输出应该作为下一层计算的输入,因此R/W SRAM由两个存储库实现,其中一个作为输入存储单元,另一个作为输出存储单元。对于下一层的计算,两组通过交换读写角色,通过切换配置来减少额外的数据传输。

权重存储器压缩后,权值矩阵的分布变得稀疏,使不同网络的存储开销减少了很多,使得所有的权值参数都可以存储在芯片上的权值SRAM中。本文采用压缩稀疏列(CSC)格式^[5]来存储稀疏权值矩阵。对于每一列,只有非零元素,相对位置索引将存储为两个向量 N_r 和 R_i 。例如,存储下面的列

[0, 0, 2, 3, 0, 0, 5, 0, 9]

作为

$N_r=[2,3,5,9]$

$R_i=[2,0,3,1]$

整个稀疏矩阵的存储包括所有的 N_r 和 R_i 每列的向量和一个列指针向量 C_p 。其中,每个元素都指向对应列的第一个非零元素。以CSC格式按列存储稀疏矩阵的情况如图5所示。

1	0	3	0	2	2	6	0
0	3	0	0	5	0	0	0
4	0	0	0	0	0	0	3
5	0	1	0	0	8	0	7

权值	1	4	5	3	3	1	2	5	2	8	6	3	7
相对行索引	0	1	0	1	0	2	0	0	0	2	0	2	0
列指针	0	3	4	6	6	8	10	11	13				

图5 以CSC格式存储稀疏矩阵

1.3 计算阵列的工作量评估

在RNN中,在每个时间步长上对新的输入和隐藏状态执行 $M \times V$ 操作。长短期记忆(LSTM)^[7]是RNN模型中广泛使用的一个单元,它提供了更复杂的隐藏单元计算,每个LSTM单元可以分解为 $8M \times V$ 操作。由于矩阵被压缩,每个矩阵的非零元素相互不同,导致并行计算时工作负载不平衡的问题。

例如:稀疏矩阵 W 的非零元素的个数 i 表示为 n_i , W 的大小 i 表示为 n 。所以,密度可以用 $d=n_i/n$ 表示。通常情况下,RNN权重矩阵的尺寸较大。即使构建一个相对较小的RNN,只有100个隐藏节点,当它被用作一个语言模型,输出标签是1000个单词时,那么权值矩阵的大小 n 将变成 10^5 。假设在压缩后,当将4个不同密度的LSTM门单元(i, f, o, c)的 $M \times V$ 操作映射到4个RCA时,可能需要使用输入向量操作的4个RCA中的非零元素为 n_1, n_2, n_3, n_4 ,分别为 $n_3 > n_2 > n_1 > n_4$,导致每个RCA的执行时间不同。然而,由于这4个门单元的结果需要同步,作为下一个操作的输入,保持每个RCA的计算同步是很重要的,否则它将增加额外的等待时间并导致资源的浪费。

2 采用可扩展电压的低功率可扩展加速器

RCA阵列计算过程的进一步优化是通过计算单元电压调节来有效利用非关键路径的时延冗余。

2.1 基于工作负载的缩放频率和电压

(1) 工作量与频率之间的关系

为了评估工作负载并调整VF,工作负载传感器得到每个稀疏权值矩阵的非零元素数,即 n_i 。设置两个阈值 W_l 和 W_h 。当工作负载非零元素的编号为 n 时 i 小于 W_l ,电

压频对应为 $V_l - F_l$ ；当 n_i 大于 W_l 但比 W_h 小时，电压频对应为 $V_m - F_m$ ；当 n_i 大于 W_h 时，电压频对应为 $V_h - F_h$ 。

(2) 基于频率的动态电压分级研究

CMOS 电路耗散的动力功率主要用下式描述

$$P_{\text{dyn}} = C_L \times VDD^2 \times f_{\text{clock}}$$

电压项对功耗的影响最大，电源电压的降低会导致功耗的急剧降低。一般来说，为了最大地降低功耗，动态电压缩放系统试图找到支持该时钟频率的最低供电电压。动态电压比例化的原理如图 6 所示。在情况 1 中，假设系统的电源电压为 V_1 ，任务必须在时间 T 之前完成，但事实是在时间 T 之前任务 1 已完成，系统将暂时处于空闲状态，直到执行下一个任务，松弛时间足够大，系统可以正常工作。在情况 2 中，系统的电源电压为 V_2 ($V_2 = 0.5V_1$)，任务在时间 t_2 时完成，定时松弛等于系统克服电压和温度变化的最小定时间隙。情况 2 的功耗约为情况 1 的 25%。

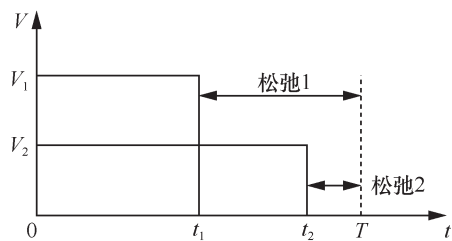


图 6 动态电压比例化的原理

动态电压调压器系统如图 7 所示。分频器使时钟频率

根据工作量除以 2 或 4。分布在处理阵列中不同位置的监视器，用于监测电压和温度的变化。假设电压降低，并且没有足够的时间空间使处理阵列正确工作，来自监视器的 Pre_error 信号将变高，电压调节器控制 DC-DC 以提高处理阵列的电压。假设处理阵列在高频工作，处理阵列中关键路径的定时松弛远大于 0，Pre_error 信号拉低，电压定期控制 DC-DC 以降低处理阵列的电压从而节省功耗。

动态电压比例化的工作过程如图 8 所示。第一步，所有处理阵列的电源电压提高到最高值，然后分频器产生新的时钟频率，电压调节器控制处理阵列的电源电压降低到将支持该时钟频率的最低电源电压。假设 Pre_error 信号变高，电压调压器通过 DC-DC 电路提高电源电压，直到 Pre_error 信号变低。

2.2 在处理阵列中采用电压可扩展性

(1) 处理阵列采用频压监视器

该监视器由关键路径和错误检测电路组成，关键路径由寄存器和逆变器组成，处理阵列中关键路径的时延最大，通过监测该关键路径可以获得处理阵列的定时松弛。根据图 9 所示的误差检测电路的基本工作原理，使用具有时延数据输入的附加触发器来检测定时误差，通过具有两个触发器输出的 XOR 门来检测定时故障。在第一阶段，数据信号在时钟信号的上升边缘之前足够早地到达，在两个触发器中都不会发生定时冲突。在下一阶段，降低电压，增加了监测路径的时延，首先在并行触发器中发生设置违规，Q 信号与 Q_{re} 信号不同，触发误差

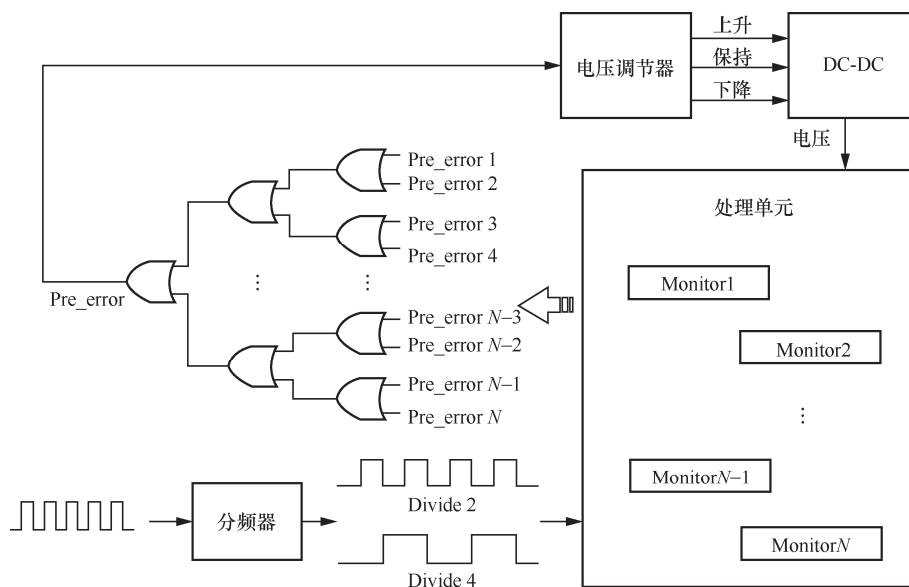


图 7 动态电压调压器系统

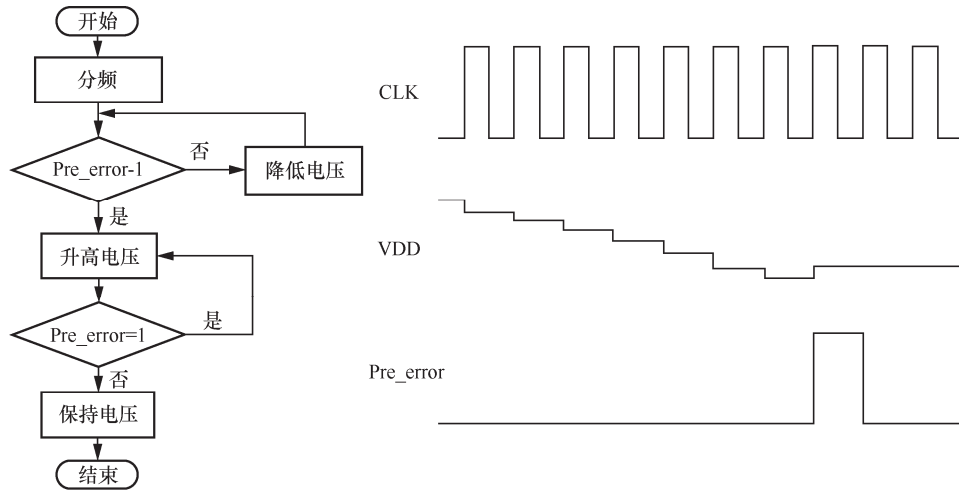


图 8 动态电压比例化的工作过程

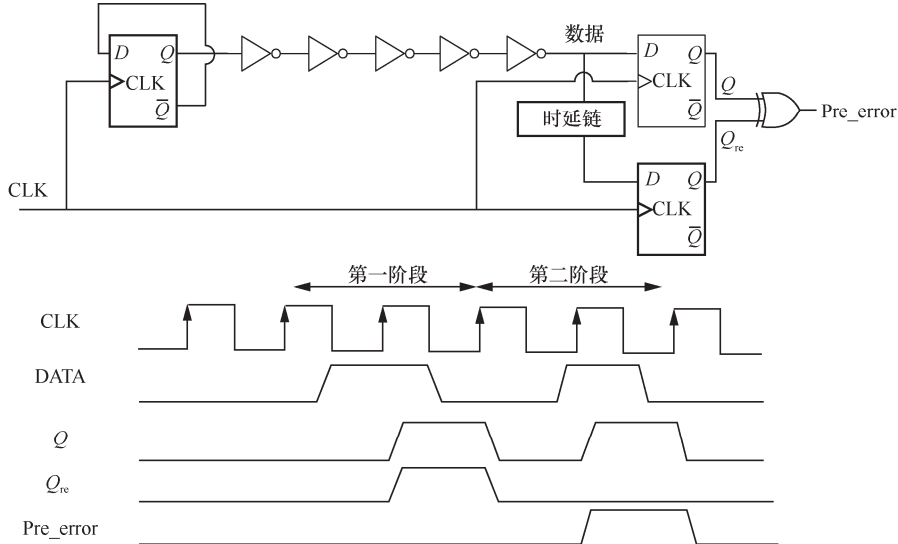


图 9 监控器的工作过程

预测信号 Pre_error ，这种情况下需要将电压增加，以确保安全定时裕度。

(2) 稀疏 RNN 低功率加速器的电压和频率

在本文的体系结构中，电压-频率缩放（VF-缩放）模块运行如图 10 所示。在工作负载传感器预测了工作负载的执行时间后，由 VF 缩放控制器将其转换为相应的频率和电压，并检查是否有空闲的数组。

步骤 1 存在空闲数组，然后将数据发送到数组，并根据工作负载设置新的电压-频率对。

步骤 2 存在缓冲队列未满足的电压匹配阵列，然后将数据发送到该阵列的缓冲队列。

步骤 3 没有电压匹配数组，然后找到最近的 VF 数组，并将数据发送到相应的缓冲区队列。

步骤 4 所有数组的输入缓冲区都已满，然后等待，直到任一数组的计算完成。

通过频压监视器和上述算法过程，可以有效提升网络在阵列上的处理效率。

3 案例研究和仿真结果

本文将所提出的电压-频率可扩展的方法应用于 ESSEN^[8]，这是一种使用 4 层双向 LSTM 模型的端到端语音识别。实验在中文语音数据库 THCHS30^[9]上进行。基于该稀疏 RNN 低功率加速器的 EESEN LSTM 网络结构在 FPGA 进行验证，使用的 FPGA 型号为 V6 XC6vlx75t-3。RNN 低功率加速器 FPGA 实现见表 1，EESEN LSTM 网络结构在 V6 XC6vlx75t-3 的布局布线图如图 11 所示。

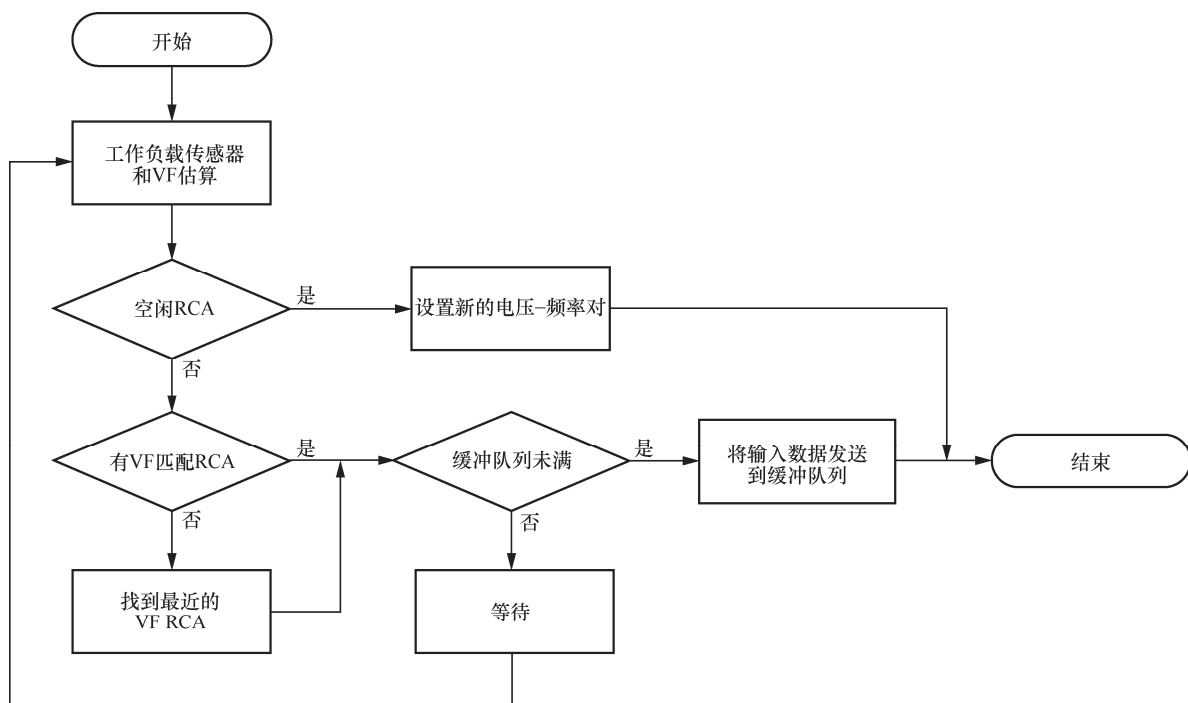


图 10 VF-缩放模块运行

表 1 RNN 低功率加速器 FPGA 实现

FPGA 型号	频率	面 (Slice)	性能/ (GOPS·s ⁻¹)	面积效率/ (MOPS·s ⁻¹ ·Slice ⁻¹)
V6 XC6vlx75t-3	270	20.15k	128	6.35

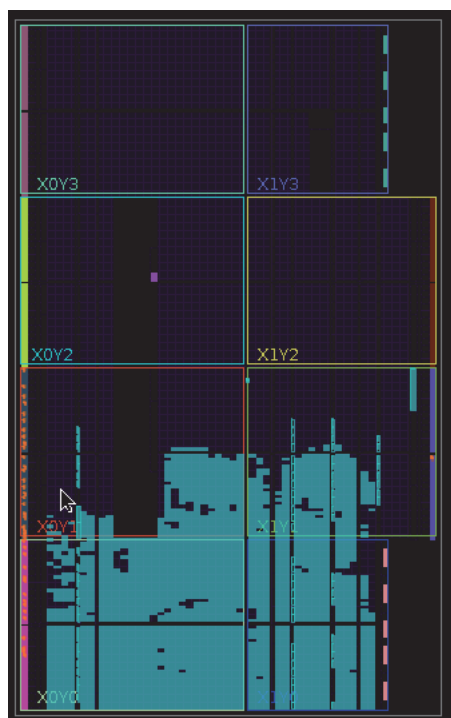


图 11 EESEN LSTM 网络结构在 V6 XC6vlx75t-3 的布局布线图

首先,对 EESEN 网络进行训练,准确率达到 92.75%。在此基础上,根据深度压缩中提供的方法对网络参数进行压缩,将参数数量从 2.25 M 减少到 64 K,准确率保持在 90%,极大地降低了权重参数的存储开销。

基于本文的体系结构,整个网络的计算遵循了第 2 节中的电压-频率调节方法。系统通过读取 CSC 格式以权重 SRAM 存储的每个稀疏矩阵的列指针来确定工作负载信息。然后将每个稀疏矩阵发送到最合适的数组中进行计算。

实验结果表明,该稀疏-RNN 低功率加速器对 EESEN LSTM 网络的处理性能为 128 GOPS/s,在 TSMC 45 nm CMOS 工艺下平均功耗仅为 224 mW。与 EIE^[3]相比,稀疏 RNN 低功率加速器显示出更高的吞吐量和能量效率,使用 2 万逻辑门的资源,性能提升 55%,功耗降低 40%。

4 结束语

本文介绍了一种低功率稀疏 RNN 加速器。它可以支持压缩 RNN 加速,通过电压-频率机制实现工作负载平衡。在 FPGA 上进行开发并在边缘设备上验证,结果表明稀疏 RNN 低功率加速器具有较高的能量效率,性能优于其他稀疏 RNN 加速器,可有效提升边缘设备的 RNN 模型处理能力,同时降低计算功耗。这对于低功耗边缘设备的智能化应用,涉及自然语言交互、图像识别、语义感知等场景的方案优化,有着积极的推进作用。

参考文献：

- [1] JITENDRA, KUMAR. Long short term memory recurrent neural network (LSTM-RNN) based workload forecasting model for cloud datacenters[J]. *Procedia Computer Science*, 2018, 125: 676-682.
- [2] RAHMAN M A, AHMED F, ALI N. Contextual deep search using long short term memory recurrent neural network[C]//*Proceedings of 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*. Piscataway: IEEE Press, 2019: 39-42.
- [3] HAN S, LIU X Y, MAO H Z, et al. EIE: efficient inference engine on compressed deep neural network[C]//*Proceedings of 2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)*. Piscataway: IEEE Press, 2016: 243-254.
- [4] HAN S, MAO H, DALLY W J. Deep compression: compressing deep neural networks with pruning, trained quantization and Huffman coding [J]. *Fiber*, 2016, 56(4):3-7.
- [5] DORRANCE R, REN F B, MARKOVIĆ D. A scalable sparse matrix-vector multiplication kernel for energy-efficient sparse-blas on FPGAs[C]//*Proceedings of the 2014 ACM/SIGDA international symposium on Field-programmable gate arrays*. New York: ACM, 2014: 161-170.
- [6] PAL S, PARK D H, FENG S Y, et al. A 7.3 M output non-zeros/J sparse matrix-matrix multiplication accelerator using memory re-configuration in 40 nm[C]//*Proceedings of 2019 Symposium on VLSI Circuits*. Piscataway: IEEE Press, 2019: 150-151.
- [7] CHAKRABORTY S, BANIK J, ADDHYA S, et al. Study of Derivation models[C]//*Proceedings of 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)*. Piscataway: IEEE Press, 2020: 1-5.
- [8] MIAO Y J, GOWAYYED M, METZE F. EESSEN: End-to-end speech recognition using deep RNN models and WFST-based decoding[C]//*Proceedings of 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. Piscataway: IEEE Press, 2015: 167-174.
- [9] WANG D, ZHANG X. THCHS-30: a free Chinese speech corpus[EB]. 2015.

[作者简介]

金磐石（1965—），男，中国建设银行股份有限公司首席信息官，主要从事信息技术系统战略策划、规划、协调和实施工作。

李俊杰（1978—），男，现就职于建信金融科技有限责任公司，主要从事人工智能推理技术研究工作。

王静逸（1990—），男，现就职于建信金融科技有限责任公司，主要从事人工智能在金融科技领域的应用研究工作。

李鹏翀（1981—），男，浪潮电子信息产业股份有限公司网络研发部总经理，主要从事数据中心架构研究工作。

邢磊（1981—），男，建信金融科技有限责任公司基础技术中心副总裁，主要从事分布式架构的设计研发工作。

李晓栋（1982—），男，中国建设银行股份有限公司金融科技部技术架构管理处副处长，主要从事技术架构设计工作。