本文引用格式：陈思浩,吴黎明,彭克锦,等.基于 ZYNQ 平台的卷积神经网络加速器设计与实现[J].自动化与信息工程,2024,45(1):30-34.

CHEN SiHao, WU LiMing, PENG KeJin, et al. Design and implementation of convolutional neural network accelerator based on ZYNQ platform[J]. Automation & Information Engineering, 2024,45(1):30-34.

基于 ZYNQ 平台的卷积神经网络加速器设计与实现

陈思浩　吴黎明　彭克锦　许志杰

（广东工业大学机电工程学院，广东 广州 510006）

摘要：针对卷积神经网络模型规模较大，以及嵌入式系统计算资源有限的问题，提出一种基于 ZYNQ 平台的卷积神经网络加速器设计方案。采用软硬件协同设计的原则，首先，在 FPGA 端设计图像、参数输入模块；然后，利用 FPGA 并行计算技术实现卷积层和池化层运算，并通过摄像头采集手写数字图像与 LCD 显示结果；最后，在嵌入式平台上实现手写数字识别。实验结果表明，卷积层和池化层的运算速度比 ARM 平台提高了 2.68 倍。

关键词：卷积神经网络；ZYNQ 平台；硬件加速；FPGA

中图分类号：TN912.3　　**文献标志码：**A　　**文章编号：**1674-2605(2024)01-0005-05

DOI：10.3969/j.issn.1674-2605.2024.01.005

Design and Implementation of Convolutional Neural Network Accelerator Based on ZYNQ Platform

CHEN Sihao　WU Liming　PENG Kejin　XU Zhijie

(School of Electromechanical Engineering, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: A convolutional neural network accelerator design scheme based on the ZYNQ platform is proposed to address the issues of large-scale convolutional neural network models and limited computing resources in embedded systems. Adopting the principle of software hardware collaborative design, first, design image and parameter input modules on the FPGA side; Then, using FPGA parallel computing technology to implement convolutional and pooling layer operations, and capturing handwritten digital images and LCD display results through a camera; Finally, implement handwritten digit recognition on an embedded platform. The experimental results show that the computational speed of the convolutional and pooling layers is 2.68 times faster than that of the ARM platform.

Keywords: convolutional neural networks; ZYNQ platform; hardware acceleration; FPGA

0　引言

随着人工智能技术的迅速发展，卷积神经网络（convolutional neural network, CNN）作为一种深度学习模型，广泛应用于图像识别[1]、目标检测[2]和语音处理[3]等领域。然而，传统的嵌入式系统无法满足 CNN 的复杂性和大规模计算的需求。图形处理器（graphics processing unit, GPU）虽然可以加速 CNN，但其存在体积大、功耗高等问题。为此，研究人员将硬件加速器应用于 CNN 的计算中，以提高其计算性能和能效。因此，在保证性能的前提下，体积更小、功耗更低的硬件平台成为 CNN 加速领域的热门发展方向[4-7]。

基于现场可编程门阵列（field-programmable gate array, FPGA）的加速平台因具有可编程性强、并行计算能力强等特点，成为研究热点[8-10]。但直接在 FPGA 上实现 CNN 计算是一项复杂的任务，需考虑诸多因素，如外设控制、内存带宽、开发难度和开发周期等。

为此，本文提出一种基于 ZYNQ 平台的卷积神经

网络加速器设计方案，在 FPGA 端设计加速器模块，通过摄像头采集手写数字图像与 LCD 显示结果，实现手写数字的识别。该方案根据软硬件协同设计的原则，利用 ZYNQ 平台上 FPGA 的并行计算能力和 ARM 的通用计算能力，对 CNN 模型中的卷积层和池化层进行 IP 核设计，提升了手写数字的识别速度。

1 系统设计

1.1 系统组成

硬件系统主要由 OV5640 摄像头、FPGA、ARM、LCD 显示屏和 DDR 存储器等组成，如图 1 所示。

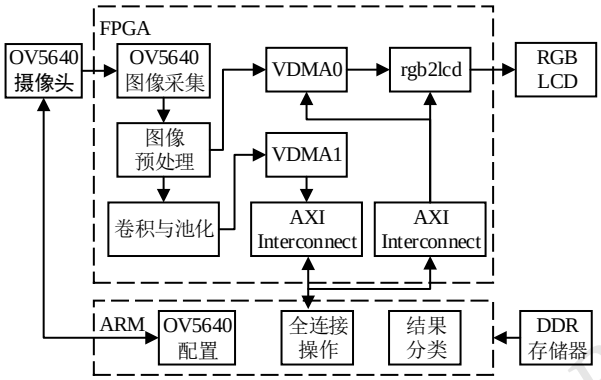

图 1 硬件系统组成框图

ZYNQ 平台是硬件系统的核心部分，它集成了 FPGA 和 ARM，通过双片 BRAM（Block RAM）与 AXI 总线实现 PS 与 PL 之间的通信，进而实现嵌入式开发[11-13]。这种组成方式允许开发人员在单个芯片上同时运行硬件计算和嵌入式软件，具有较强的灵活性，可满足不同应用场景对计算资源和实时性的需求。

硬件系统运行流程如下：

1） 在 ARM 端对 OV5640 摄像头进行配置，通过 OV5640 摄像头采集手写数字图像；

2） 手写数字图像传入 FPGA 端的 OV5640 图像采集 IP 核，并将 8 位图像数据拼接为 24 位图像数据；

3） FPGA 端的图像预处理 IP 核对 24 位图像数据进行灰度和二值化处理；

4）FPGA 端的卷积与池化 IP 核提取手写数字图像特征后，通过 AXI 总线将池化后的数据传入 ARM，进行全连接运算与结果分类；

5） 分类结果显示在 LCD 上。

1.2 CNN 结构

CNN 利用了局部连接和参数重用的特性，其每层都单独使用一组卷积核，有助于从局部相关数据中提取有用的特征[14-17]。CNN 主要包括输入层、卷积层、池化层、全连接层等，结构如图 2 所示。

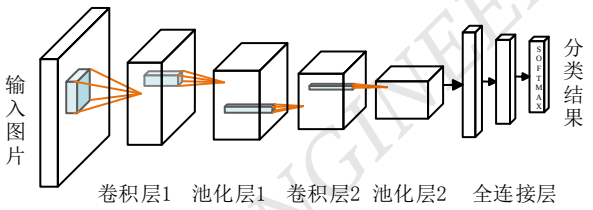

图 2 CNN 结构

考虑到嵌入式芯片的计算资源有限，为充分发挥 FPGA 端和 ARM 端的性能，本文对经典的 CNN 结构进行改进，在尽量精简结构的同时，保留了 CNN 的卷积层。改进后的 CNN 结构包含 1 个卷积层、1 个池化层和 2 个全连接层，以实现手写数字的识别。输出层有 10 个节点，每个节点对应 1 个手写数字，因此改进后的 CNN 结构没有使用 SoftMax 函数。如果需要部署更复杂的 CNN，只需在加速器模块中导入新的权重参数，并复用卷积层和池化层 IP 核即可。

2 加速器模块设计

加速器模块主要由图像输入模块、参数输入模块、卷积运算模块和池化运算模块组成，结构框图如图 3 所示。

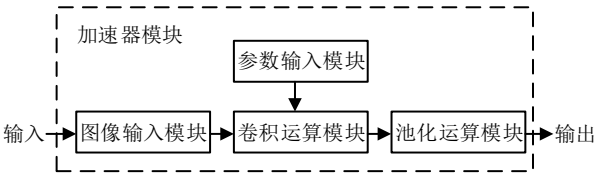

图 3 加速器模块结构框图

在 FPGA 上部署神经网络一般采用硬件描述语言（hardware description language, HDL）和高层次综合（high-level synthesis, HLS）工具两种方法。虽然传统的 HDL 编程耗时比 HLS 长，但它可以精确定义每

个时序硬件电路的行为和功能，能更好地利用 FPGA 资源。因此，本文采用 HDL 设计加速器模块。

2.1 图像输入模块

图像输入模块主要由降采样部分和数据存储单元组成。考虑到采集的手写数字图像需清晰地显示在 800×600 像素的 LCD 上，在输入卷积运算模块前，需对其进行降采样操作。

根据卷积层的输入大小，先对采集的手写数字图像进行倍数放大，卷积层输入图像的大小为 28×28 像素，将手写数字图像先放大 4 倍，即图像大小为 112×112 像素；再进行降采样操作，即每隔 4 个点取 1 个点，如图 4 所示。

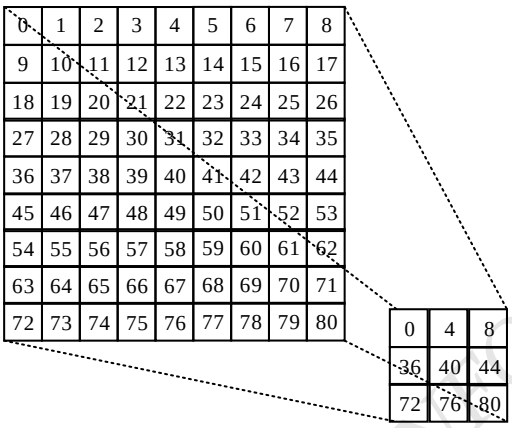

图 4 降采样示意图

数据存储单元采用 BRAM 来实现。本文设计一个 28 bit 的 BRAM 来存储手写数字图像数据。降采样后的手写数字图像数据以列优先的顺序写入 BRAM，每次写入 1 个像素点的数据，即 1 bit。

读端口和写端口通过不同的使能信号控制。读端口的使能信号一直为高电平，可连续从 BRAM 中读取卷积窗口大小的数据。读取数据时，一次性读取一行数据的前 5 个数据，每次读取后，数据指针向后移动 1 位，这样可确保连续读取 5×5 的数据，满足卷积计算的需求，如图 5 所示。

写入数据时，通过行计数器控制写使能信号，每采集 1 个像素点，就向 BRAM 写入 1 位数据。当 BRAM 存储完整图像的一行数据后，再切换至下一行。

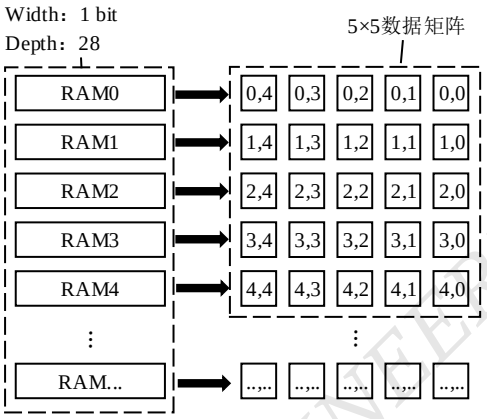

图 5 图像数据读取图

读写两个端口使用不同的地址进行控制，不仅能避免读写冲突，还能够高效地实时读取 BRAM 中卷积窗口大小的数据，为后续的卷积运算提供数据支持。

2.2 参数输入模块

参数输入模块使用 6 个 ROM 来存储卷积层中每个卷积核 5×5 窗口内不同位置的参数值和偏置值，其存储方式与图像输入模块存储图像数据相似。

2.3 卷积运算模块

卷积运算模块作为实现神经网络前向传播的核心模块，利用一维卷积来计算图像与每个卷积核对应的响应值。其中，每个卷积核对应一个卷积窗口的权重参数。卷积运算模块中包含多个乘累加器，如图 6 所示。

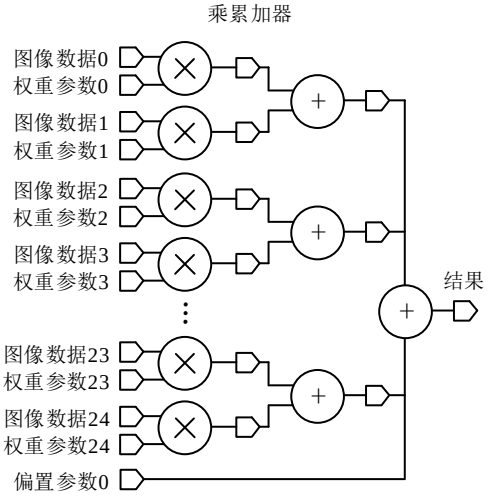

图 6 卷积运算图

控制单元根据行计数和卷积核索引来控制乘累加器的操作。每个乘累加器负责一个卷积核与图像数据矩阵的乘加计算。首先，乘累加器从BRAM和ROM中同步读取 1 个图像数据矩阵和相应的卷积核参数；然后，依次将对应的图像数据与卷积核参数相乘；最后，将结果相加。一个完整的卷积计算需要连续读取 5 行的图像数据与一个卷积核的 5×5 参数，共进行 25 次乘加运算。为提高计算效率，每个乘累加器将 25 次乘加运算并行化为 5 次流水线操作，充分利用每个时钟周期的计算资源，提升了卷积层的计算吞吐量。

2.4 池化运算模块

池化运算模块采用2×2的池化窗口对每个2×2块数据进行最大值池化。池化运算模块主要包括比较器和 FIFO 两部分，计算过程如图 7 所示。

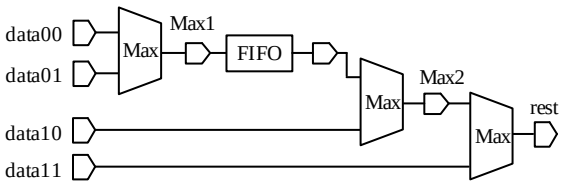

图 7 池化运算图

比较器接收来自卷积运算模块的 2×2 块数据，每次比较该块内第一组的两点数据，输出最大值，存入 FIFO 中。当第二组 2×2 块数据进入比较器时，将该块内的值与 FIFO 保存的最大值进行比较，更新最大值。每处理完一组 2×2 块数据后，输出的最大值即为该组数据池化的结果。

为获得池化结果的正确顺序，需控制比较器和 FIFO 的读写时序。比较器每完成一次 2×2 块数据的最大值计算后，将结果立即写入 FIFO，同时 FIFO 读端口被使能输出结果。读写两端口在不同的时钟边沿分别工作，保证数据的有序输出。最后，通过串联多个 2×2 块数据的最大值计算，实现整个输入特征图的最大池化。

池化后的低维特征图作为 ARM 后端程序的输入，经过 VDMA 传输到 ARM，提供给全连接层进行计算。

3 实验分析与结果

本实验采用领航者 ZYNQ 开发板，其主芯片 ZYNQ 采用 XC7Z020CLG400-2，ARM 端采用频率为 666 MHz 的双核 Cortex-A9 处理器，FPGA 端的时钟频率为200 MHz，开发环境为Vivado2020和Vitis2020。ZYNQ 开发板带有摄像头模块接口和 RGB LCD 接口。考虑到适合手写数字识别的应用场景，开发了摄像头和 LCD 等模块来模拟真实识别场景。上板验证效果如图 8 所示。

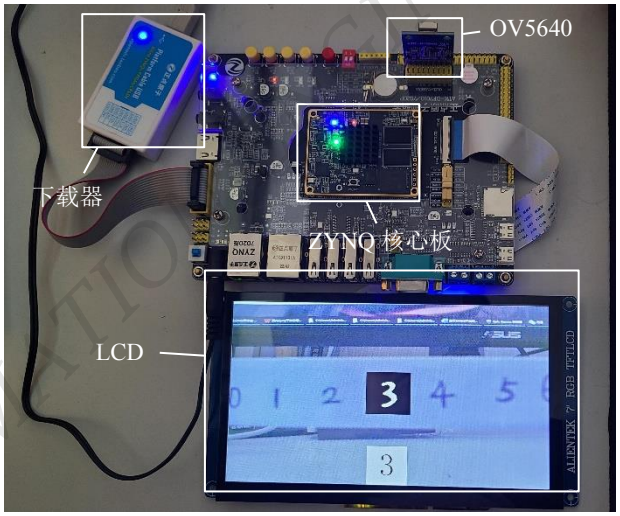

图 8 上板验证效果图

在开发工具 Vivado2020 中，生成器件的资源使用情况报告如表 1 所示。

表 1 硬件平台资源使用情况

资源	总资源数/个	使用资源数/个	使用率%
LUT	53 200	11 894	22.36
LUTRAM	17 400	360	2.07
FF	106 400	14 207	13.35
BRAM	140	35	25.00
IO	125	45	36.00
BUFG	32	4	12.50
MMCM	4	1	25.00

通过多次实验分析加速器模块的仿真时序图，计算卷积与池化运行所需的时间，在 Vitis 中通过时间获取函数得到全连接层的运算时间，即可得到 FPGA 加速后的 CNN 运行总时间，并将其与仅在 ARM 端

运行的 CNN 进行对比，结果如表 2 所示。

表 2　ARM 与 FPGA+ARM 平台的 CNN 运行时间

平台	识别精度/%	运行时间/ms
ARM	98.8	64.18
FPGA+ARM	98.6	23.94

由表 2 可知，相较于仅在 ARM 端运行的 CNN，FPGA 加速后的 CNN 在识别精度损失较小的情况下，网络运行时间减少了 2.68 倍。

4　结论

本文基于 ZYNQ 平台提出了一种卷积神经网络加速器设计方案。在 FPGA 端设计了图像数据与参数数据存储模块，实现高效的存储与读取，为卷积计算提供数据支持。采用并行设计的思路实现卷积和最大池化的运算，在保证识别精确度的同时，卷积层和池化层的运行速度提高了 2.68 倍。与其他神经网络的加速方案相比，该加速方案具有功耗低、体积小、容易部署、通用性强等特点，具有一定的实际应用意义。

参考文献

[1] NARAYAN A, MUTHALAGU R. Image character recognition using convolutional neural networks[C]//2021 Seventh International conference on Bio Signals, Images, and Instrumentation (ICBSII). IEEE, 2021:1-5.

[2] YAN X, SHUAI C, ZHENG H. A Yolov3-based multi-target detection system for complex scenes[C]//2021 2nd International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT). IEEE, 2021:327-332.

[3] HSU Y, LEE Y, BAI M R. Learning-based personal speech enhancement for teleconferencing by exploiting spatial-spectral features[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022:8787-8791.

[4] USHIROYAMA A, WATANABE M, WATANABE N, et al. Convolutional neural network implementations using Vitis AI [C]//2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC). IEEE, 2022: 0365-0371.

[5] ADIONO T, SUTISNA N. FPGA based hardware accelerator design for convolution process in convolutional neural network [C]//2021 International Conference on Electrical Engineering and Informatics (ICEEI). IEEE, 2021:1-5.

[6] XIONG Z M. A survey of FPGA based on graph convolutional neural network accelerator[C]//2020 International Conference on Computer Engineering and Intelligent Control (ICCEIC). IEEE, 2020:92-96.

[7] LI L, CHEN X, GAO W. Implementation of convolutional neural network accelerator based on ZYNQ[C]//2022 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA). IEEE, 2022:158-165.

[8] PISHARODY J N, PRANAV K B, RANJITHA M, et al. FPGA implementation and acceleration of convolutional neural networks[C]//2021 6th International Conference for Convergence in Technology (I2CT). IEEE, 2021:1-4.

[9] 江瑜,朱铁柱,蒋青松,等.基于 FPGA 的卷积神经网络硬件加速器设计[J].电子器件,2023,46(4):973-977.

[10] 黄沛昱,赵强,李煜龙.基于 FPGA 的卷积神经网络硬件加速器设计[J].计算机应用与软件,2023,40(3):38-44.

[11] 冯光顺,应三丛.ZYNQ 的卷积神经网络硬件加速通用平台设计[J].单片机与嵌入式系统应用,2019,19(3):3-6;9.

[12] 刘晛,吴瑞琦,高尚尚,等.基于 ZYNQ 的通用型卷积神经网络设计与实现[J].电子器件,2023,46(1):121-125.

[13] 缪丹丹,张鹏,张鑫宇,等.基于 ZYNQ 平台的通用卷积加速器设计[J].国外电子测量技术,2022,41(11):72-77.

[14] 季长清,高志勇,秦静,等.基于卷积神经网络的图像分类算法综述[J].计算机应用,2022,42(4):1044-1049.

[15] 谭亚红,史耀.完备变分模态分解和多传感器卷积神经网络的轴承故障诊断方法[J].机床与液压,2022,50(14):182-188.

[16] 刘斌,龙健宁,程方毅,等.基于卷积神经网络的物流货物图像分类研究[J].机电工程技术,2021,50(12):79-82;175.

[17] 许富景,陈长颖,杜少成.基于改进 CNN 的压缩感知自然图像重建方法[J].中国测试,2022,48(9):7-16.

作者简介：

陈思浩，男，1998 年生，在读研究生，主要研究方向：智能测控。E-mail: edwardchenx@foxmail.com

吴黎明，男，1962 年生，硕士，教授，主要研究方向：智能测控。E-mail: jkyjs@gdut.edu.cn

彭克锦，男，1999 年生，在读研究生，主要研究方向：智能测控。E-mail: 1334152998@qq.com

许志杰，男，1999 年生，在读研究生，主要研究方向：智能测控。E-mail: 1422411797@qq.com