基于 PYNQ 的手写体数字识别系统设计实现

谷文成,高谷九祥,凌卓毅,肖　建
(南京邮电大学 电子与光学工程学院、微电子学院,江苏 南京 210023)

摘　要:手写体数字的识别作为图像识别领域的一项重要分支,广泛应用于银行汇款单号、人口普查、财务报表等大规模数据统计领域中。而传统的手写体数字识别系统一般采用 CPU 或 GPU 的平台,有着功耗和成本较高、难以部署在移动端等弊端。针对上述问题,设计了一种以卷积神经网络为基础,基于 PYNQ 的手写体数字识别系统。通过软硬件协同设计的方式,合理划分软硬件任务来降低系统功耗。首先在电脑端搭建卷积神经网络模型,通过训练验证,以获取权重和偏置等技术参数,并转换为相应的二进制格式文件。之后在 VIVADO HLS 工具中设计完成了卷积层和最大池化层的 IP 核模块设计,以及系统的连线。最后设计实现对应的上位机程序进行调控。在 MNIST 数据集的测试下,该系统的识别准确率达到了 99.07%,功耗仅为 1.54 W,相比于其他类似工作具有明显优势。

关键词:手写体数字识别;卷积神经网络;PYNQ;软硬件协同设计;移动端部署
中图分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1673-629X(2022)0031-05

Design and Implementation of Handwritten Digit Recognition System on Mobile Terminal Based on PYNQ

GU Wen-cheng, GAO Gu-jiu-xiang, LING Zhuo-yi, XIAO Jian
(School of Electronic and Optical Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: As an important branch in the field of image recognition, the recognition of handwritten digits is widely used in large-scale data statistics fields such as bank remittance number, census, and financial statements. However, traditional handwritten digit recognition systems are generally based on CPU or GPU platforms, which have disadvantages such as high power consumption and cost, and difficulty in deploying on mobile terminals. In response to these problems, we design a PYNQ-based handwritten digit recognition system based on convolutional neural networks. Through the idea of software and hardware co-design, tasks are divided reasonably. Firstly, the convolutional neural network model is built on the PC side. Through training and verification, the technical parameters such as weight and bias are obtained and converted into the corresponding binary format file. Afterwards, the IP core modules of the convolutional layer and the maximum pooling layer were designed in the VIVADO HLS tool, as well as the connection of the system. Finally, the corresponding host computer program is designed and implemented for regulation. Under the test of the MNIST data set, the recognition accuracy of the system reached 99.07%, and the power consumption was only 1.54 W, which has obvious advantages compared to other similar tasks.

Key words: handwritten digit recognition; convolutional neural network; PYNQ; software and hardware co-design; mobile terminal deployment

0　引　言

目前,手写体数字识别技术作为人工智能领域的一项重要标杆研究,在邮政、交通、金融、教育等行业的实践活动中有着广泛应用[1]。然而以 OCR、SVM、Boosting 等为代表的传统图像识别技术需要人工来完成特征信息的定义,而且对设备的要求很高。这存在着识别效果不佳、功耗较高等弊端[2],在面对灵活易变的实际应用场景时显得力不从心,因此高效准确且易用的手写体数字识别系统的研究刻不容缓。

为了适应手写体数字识别高精确度的特点,目前卷积神经网络(CNN)已经成为主要驱动力。它可以快速准确地发掘数据局部特征,并在短时间内完成全局训练特征提取[3]。相较于人工定义的特征信息,通过对神经网络训练得到的特征参数及其模型更为合

收稿日期:2021-10-12
基金项目:江苏省大学生创新训练计划(SZDG2019007);南京邮电大学国自孵化项目(NY220013)
作者简介:谷文成(1999-),男,研究方向为智能信息处理;通讯作者:肖　建(1976-),男,博士,教授,从事嵌入式系统应用研究与学生创新、实验室建设与管理等工作。

理,因此其对于图像特征的识别也更为准确。而卷积神经网络作为一种网络结构,有着高并行性的特点,对硬件平台具有一定的要求。专门的图形处理器(GPU)虽然能够解决这一问题,但其功耗往往偏高。而拥有更大规模的逻辑和计算单元的现场可编程逻辑门阵列(FPGA)在性能、并行运算、功耗和尺寸等方面都具备了明显优势[4]。

该文采用 PYNQ-Z2 开发平台来完成系统设计,其以 PYNQ 的 PS(Process System)端作为核心来合理分配软硬件任务[5],通过 PL(Programmable Logic)端来实现卷积神经网络中图像数据的计算,以期达到手写体识别的高准确度和降低功耗的目的。

1 系统总体设计

系统采用 Xilinx 公司的 PYNQ-Z2 型号的 FPGA 作为开发平台[6],同时选择通过软硬件协同的方式来进行设计。PYNQ-Z2 开发板包括 PS 和 PL 两个部分,在 PS 端,包含一个采用双 ARM Cortex A9 内核的处理系统,其中包含了指令、数据缓存和内存单元等大量片上缓存资源;而 PL 端,则是传统的 FPGA 结构,具备大量的可编程逻辑和高效的计算能力,有极强的可扩展性。

因此,根据以上二者的特点,对软硬件的功能进行划分。PL 端负责 CNN 中核心的卷积运算和最大池化层的密集计算部分,起到并行、加速的效果,并降低功耗;PS 端承担上位机的角色,编写驱动程序,对底层进行控制,具有更强的灵活性。

系统的整体流程示意图如图 1 所示。

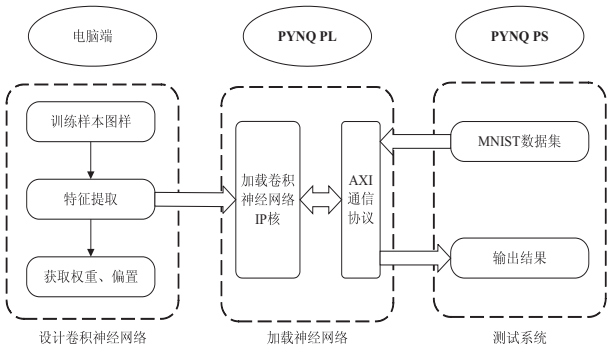


图 1 手写体数字图像识别系统构成示意图

首先在电脑端构建卷积神经网络模型,对训练样本图样进行训练,以得到权重和偏置参数。之后利用 VIVADO HLS 工具在 PYNQ 的 PL 端进行卷积层和最大池化层的 IP 核模块设计,并构建完整的系统电路模块。最后,通过起到主控作用的 PS 端对测试数据集进行读取分析,并通过 AXI 协议与 PL 端进行数据传输[7]。PL 端进行运算识别后,将结果反馈给 PS 端,并输出识别结果。

2 卷积神经网络的设计与实现

2.1 卷积神经网络的基本结构

卷积神经网络结构包括:输入层(Input layer)、卷积层(Convolutional layer)、池化层(Convolutional layer)和全连接层(Fully-Connected layer)[8]。每一层有多个特征图,每个特征图上都被自定义了一种卷积核,通过卷积核可以实现对上一层图像的特征进行提取,每个特征图都储存着上一层特征图被卷积计算后的特征信息[9]。其中,输入层通常用于接受输入图像,而这些输入图像往往通过灰度处理来适应卷积层开展工作的需要;卷积层最重要的作用是实现特征的提取功能,即通过卷积核与原始图像相乘相和,因此卷积层是 CNN 的核心图像处理模块,也是大部分计算发生的地方;池化层的作用是进行进一步特征抽样,并实现数据降维,减少输入中的参数数量;全连接层主要用于集合所提取到的特征,并通过对比给出识别的结果。

2.2 卷积神经网络设计流程

手写体数字图像识别系统需要在电脑端完成卷积神经网络模型的设计与训练,最终通过对数据集的处理来得到特征参数相应的二进制文件。该文件用于卷积神经网络中全连接层中权重(weight)和偏置(bias)的配置,从而判断数据集中图像上的数字[10]。

MNIST 数据集被广泛用作不同机器学习和模式识别建议的测试平台。其共包含 70 000 个实例,其中 60 000 个用于训练特征参数,其余用于测试系统[11]。首先,原始图像被提交进行预处理。此过程首先涉及将图像标准化以适合 20×20 像素框,同时保留纵横比。然后,应用抗锯齿滤波器,结果黑白图像有效地转换为灰度。最后,引入空白填充以将图像放入更大的 28×28 像素框,以便数字的质心与其中心匹配。对 MNIST 数据集进行识别的卷积神经网络的结构模型如图 2 所示。输入的手写体数字图像经过卷积层和池化层的处理,卷积神经网络捕获其特征。随后将这些特征传递到全连接层,在权重和偏置的处理之下,通过对比获得识别的结果,并将其输出。

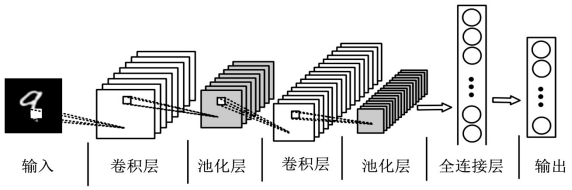


图 2 卷积神经网络的结构模型

2.3 VIVADO IP 核模块设计

完成卷积神经网络的研究与设计之后,在电脑端

搭建卷积神经网络的模型并训练数据集。在得到特征参数二进制文件后,便需要在电脑端完成卷积层以及池化层 IP 核模块的设计。主要使用了 VIVADO HLS 开发工具[12],其中 C 语言将功能描述转换为硬件描述语言,用于硬件综合。HLS 的设计流程一般可以分为三个部分,主要是设计系统输入、对系统代码进行仿真和获得系统输出[13]。

2.3.1　卷积层 IP 核模块设计

图 3 给出了卷积结构的经典实现,其运算需要 6 层循环[14]。需要将其在 VIVADO HLS 工具中进行实现。在遵循 HLS 编码规范的前提下,添加相应的头文件和 C 源文件。

```
for j = 1:Nom                #循环 1:遍历每个输出通道
  for x = 1:Nox              #循环 2:输出通道中的每一行
    for y = 1:Noy            #循环 3:对于行中的每个对象
      for i = 1:Nim          #循环 4:在所有输入通道上进行卷积
        for kx = 1:K         #循环 5:核的 x 维
          for ky = 1:K       #循环 6:核的 y 维
            out(j, x, y) += in(i, x + kx, y + ky) * weight
(j, i, kx, ky)
          end
        end
      end
      #加上输出偏置
      out(j, x, y) += bias(j)
    end
  end
end
```

图 3　卷积结构的经典实现

在 C 代码中,要设计卷积层的实现,首先定义卷积 IP 核的顶层函数 Conv。该函数包括输入特征、输出特征、权值特征和偏置量特征四个数组格式的特征参数。接着需进行约束。本次设计采用 AXI Lite 总线协议,约束命令如下:

#pragma HLS INTERFACE s_axilite port=INPUT

#pragma HLS INTERFACE s_axilite port=OUTPUT

接着考虑 PS 端。PS 作为上位机的角色,承担着对数据的传输工作,这意味着其必须能主动地读取参数文件。因此,mode 参数必须设置为 m_axi;同时,PS 端从属于 AXI Lite 协议的接口,因此 offset 设置为 slave。

卷积过程中的不同通道的数据之间互相独立,因此可以利用这一点提供并行度。常用的方法有 for 循环展开等。但是值得注意的是,展开的层数越多,板载资源的消耗就会越大,类似于“空间换时间”的思想。因此,需要在循环展开数和板载资源的消耗上取得一个平衡点。针对于 PYNQ-Z2 开发板,本设计将循环展开为 3 个并行的子循环,以此将性能平衡至最优。约束命令为:

#pragma HLS UNROLL factor=3

完成 C 代码的编写后,检查无误,开始启动仿真、编译并执行程序,验证没有报错后进行算法综合与 C/RTL 协同仿真,综合结束后生成性能报告。使用 HLS 对设计好的代码进行打包,并将其封装成 IP 模块,导出实验所需要的 IP 核模块。该文所设计的卷积层 IP 核如图 4 所示。

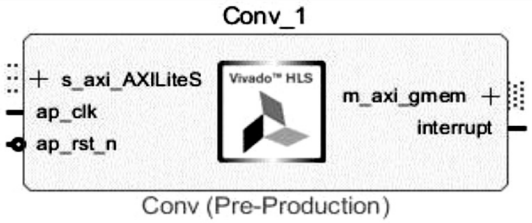


图 4　卷积层 IP 核通用电路

2.3.2　池化层 IP 核模块设计

最大池化层是对函数采样区域内的特征点取最大值,对输入的图像数据进行下采样。与卷积层的设计类似,最大池化层由 5 层循环嵌套组成。最大池化层函数的定义也与卷积层类似,参数分为输入图像长度、输入图像宽度、输入通道数、窗口长度和窗口宽度,其中输入输出特征参数为多维矩阵。

3　基于 PYNQ 的手写体图像识别系统实现

VIVADO HLS 生成的 RTL 代码需要编译到 FPGA 的比特流文件中,通过 VIVADO 设计来实现这一步骤。在 VIVADO 中导入在 HLS 中设计的“IP Catalog”文件后 Design Suite IP Catalog,可以新建一个 Block Design,VHLS 中的设计可以作为新组件添加在其中。此外,“Zynq7 处理系统”组件被添加在设计模块中并完成配置,随后运行自动连接模块完成组件自动化连接。最终得到的系统整体模块图如图 5 所示。其中,Conv_0 和 Pool_0 是在 HLS 中设计的卷积层和池化层的 IP 核,而 ZYNQ7 Processing System 模块由 Xilinx 公司设计,其正是搭载在 PYNQ 上的系统,系统电路连接由 VIVADO 自动化设计完成。

经过 Vivado Design 功能调用 IP 核,然后自动生成必要的 VHDL 或 Verilog 封装器文件并实例化 VHLS 设计,便实现了由语言到模块再到可读取数据的转变。所有时间限制和资源分配由 Vivado HLS 来设置,通过 Vivado Design Suite 中参数的合理选择,便能完成综合和实现多个不同的预设方案,并且可以顺利运行。当 RTL 综合完成时,包含二进制 FPGA 配置的比特流可以导出为(.bit)文件并加载到 PYNQ 开发板。

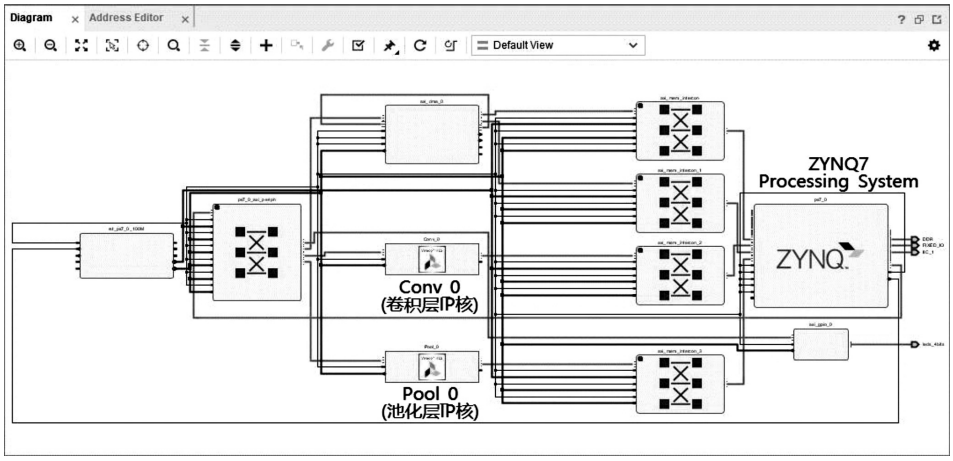


图 5 系统整体模块

4 系统测试

手写体数字图像识别系统使用 Xilinx 的 PYNQ-Z2 硬件开发板,在进行实验之前需要对本实验所使用到的开发板进行配置,主要在于开发板内部 Python 开源库的安装,以及硬件环境的配置和开发板上端口的设置。之后将 VIVADO 中生成好的比特流文件上传到开发板上,便可以使用 Python 语言对其进行编程与调用。

4.1 功耗测试

根据 VIVADO 生成的报告,可以得到 PL 端的功耗情况,如图 6 所示。可以看到,系统动态功耗为 1.402 W,静态功耗为 0.138 W,总功耗为 1.54 W。其中系统动态功耗约占总功耗的 91%。这表明,所设计的电路系统满足移动终端的低功耗需求,具有实际应用价值。

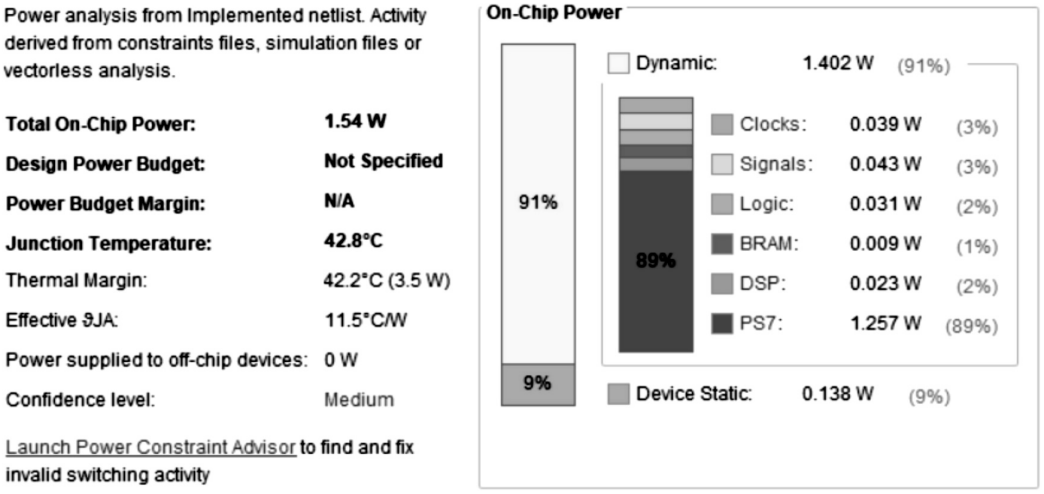


图 6 MNIST 数据集识别效果

4.2 功能测试

本系统以 MNIST 数据集为模型进行测试。MNIST 数据集是目前机器学习领域中运用最为广泛的手写体数字数据集之一,是非常具有代表性的测试对象。图 7(a)展示了对 MNIST 数据集的识别结果。图中白色背景黑色字体的为待测数据集,顶部左侧数字为手写体数字输入图像经系统识别后的输出结果,而右侧方括号中的数字为图片对应的实际分类标签,绿色字体代表系统识别图像的结果与实际分类标签结果相同,即识别手写体数字正确,红色字体代表系统识别得到的结果错误。系统电路连接实物如图 7(b)所示。

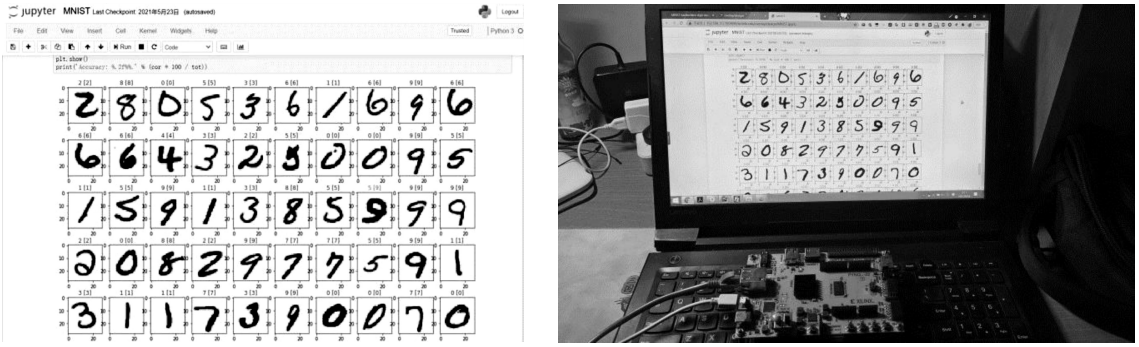
本系统以 MNIST 数据集为模型对系统进行性能测试,验证本文设计,以及与电脑端 CPU 和 GPU 性能及功耗进行比较分析。测试结果如表 1 所示[15]。

表 1 系统性能与其他平台对比

实验平台	MNIST 准确率/%	平台功耗/W	平台成本/元
Intel i5 7500	99.02	65	1 100
Nvidia GTX 1050Ti	99.02	90	1 500
PYNQ-Z2	99.07	1.54	1 000

从表 1 中可以看出,通过该系统实现对输入手写体数字图像识别,识别准确率高达 99.07%。同时,本

系统在平台功耗和平台成本方面有着巨大优势。测试结果表明提出的手写体数字识别系统具备精准、低功耗的特点,足以胜任移动端平台。且 FPGA 成本相对较低,具有良好的应用前景。



(a)MNIST 数据集识别效果　　(b)系统电路连接实物图

图 7　MNIST 数据集下系统测试图

5　结束语

基于卷积神经网络提出了一种实现手写体数字识别的方案。该方案以 PYNQ-Z2 为设计平台,基于软硬件协同设计方案,通过让硬件平台上的软硬件任务的分配合理化,最终实现了高效准确的手写体数字的识别。在 MNIST 数据集的测试下,系统的识别准确率达到了 99.07%。本系统相较于其他同类型的系统,在保持高识别准确率的情况下,还维持了较低的功耗,可以胜任部署在移动端的场景;并且 FPGA 相较于 CPU、GPU 等运算平台成本更低,在涉及图像处理的诸多领域都具备很好的应用前景。

参考文献:

[1] 陈　岩,李洋洋,余　乐,等. 基于卷积神经网络的手写体数字识别系统[J]. 微电子学与计算机,2018,35(2):71-74.

[2] 李　琼,陈　利,王维虎. 基于 SVM 的手写体数字快速识别方法研究[J]. 计算机技术与发展,2014,24(2):205-208.

[3] 章　琳,袁非牛,张文睿,等. 全卷积神经网络研究综述[J]. 计算机工程与应用,2020,56(1):25-37.

[4] 许永全,冯玉田. 基于 FPGA 的卷积神经网络动态加载 SOC 设计[J]. 计算机技术与发展,2020,30(7):1-5.

[5] 王春林. 基于 ZYNQ 的卷积神经网络软硬件协同设计研究与实现[D]. 大连:大连海事大学,2020.

[6] KÄSTNER F,JANBEN B,KAUTZ F,et al. Hardware/software codesign for convolutional neural networks exploiting dynamic partial reconfiguration on PYNQ[C]//2018 IEEE international parallel and distributed processing symposium workshops (IPDPSW). Vancouver, BC, Canada: IEEE, 2018:154-161.

[7] RANDLES B M,PASQUETTO I V,GOLSHAN M S,et al. Using the Jupyter notebook as a tool for open science:an empirical study[C]//2017 ACM/IEEE joint conference on digital libraries (JCDL). Toronto,ON,Canada:IEEE,2017:1-2.

[8] ALBAWI S,MOHAMMED T A,AL-ZAWI S. Understanding of a convolutional neural network[C]//2017 international conference on engineering and technology (ICET). Antalya,Turkey:IEEE,2017:1-6.

[9] 邹翔熙. 基于 FPGA 的卷积神经网络的 IP 化设计与实现[D]. 海口:海南大学,2020.

[10] 张素智,吴玉红,常　俊. 基于改进 AlexNet 卷积神经网络的轮胎图像识别[J]. 计算机技术与发展,2021,31(7):182-186.

[11] DENG L. The mnist database of handwritten digit images for machine learning research [best of the web][J\]. IEEE Signal Processing Magazine,2012,29(6):141-142.

[12] CORTES A,VELEZ I,IRIZAR A. High level synthesis using Vivado HLS for Zynq SoC:image processing case studies[C]//2016 conference on design of circuits and integrated systems (DCIS). Granada,Spain:IEEE,2016:1-6.

[13] GUREL M. A comparative study between rtl and hls for image processing applications with fpgas[M]. San Diego:University of California,2016.

[14] 王小雪. 基于 FPGA 的卷积神经网络手写数字识别系统的实现[D]. 北京:北京理工大学,2016.

[15] 童耀宗. 基于 FPGA 的卷积神经网络加速器的设计与实现[D]. 成都:电子科技大学,2019.