基于 PYNQ 的图像分类识别技术研究与实现

陈 禹[1],谷文成[1],渠吉庆[1],蒋志鹏[1],张 瑛[1],孙科学[1,2]*
(1. 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;
2. 射频集成与微组装技术国家地方联合工程实验室,江苏 南京 210023)

摘 要:为了实现低功耗的图像分类识别系统,设计一种基于卷积神经网络的图像分类识别系统方案,该方案研究基于 ARM+FPGA 异构系统的实现方法,系统搭载于 Xilinx 的 PYNQ 嵌入式开发平台。在电脑端对待测试的数据集搭建卷积神经网络模型并完成 MNIST 和 CIFAR-10 数据集的训练验证。随后设计特征参数提取函数完成权重和偏执参数的提取及格式转换,转换为硬件平台可以进行读取的二进制格式。接着使用 Xilinx VIVADO HLS 设计工具,设计实现图像分类识别系统中卷积神经网络的自定义 IP 核模块。完成自定义 IP 核的设计之后,以 IP 核模块和 ZYNQ 模块为主实现整体系统的通路搭建,完成验证后在 Jupyter Notebook 中通过上位机程序调用控制。最后,完成驱动程序及系统上位机的设计。测试结果表明,系统对 MNIST 和 CIFAR-10 数据集的识别可以实现分类,系统功耗仅为 1.54 W。该系统具有通用性好、硬件功耗低等优点,可广泛应用于边缘计算环境中。

关键词:卷积神经网络;软硬件协同设计;PYNQ;VIVADO;Jupyter Notebook
中图分类号:TP391.4 **文献标识码**:A **文章编号**:1673-629X(2021)12-0073-05
doi:10.3969/j.issn.1673-629X.2021.12.013

Research and Implementation of Image Classification and Recognition Technology Based on PYNQ

CHEN Yu[1], GU Wen-cheng[1], QU Ji-qing[1], JIANG Zhi-peng[1], ZHANG Ying[1], SUN Ke-xue[1,2]*
(1. School of Electronic and Optical Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210023, China;
2. Nation-Local Joint Project Engineering Lab of RF Integration & Micropackage, Nanjing 210023, China)

Abstract: In order to realize an image classification and recognition system with low power consumption, we propose an image classification and recognition system solution based on convolutional neural network, which studies the implementation method of heterogeneous system based on ARM+FPGA, and the system is mounted on Xilinx PYNQ embedded development platform. After building the convolutional neural network model on the computer-side for the tested data sets, the training and verification of MNIST and CIFAR-10 data sets are completed. After that, the feature parameter extraction function is designed to complete the weight and paranoid parameter extraction and format conversion, which is converted to a binary format that can be read by the hardware platform. Then Xilinx VIVADO HLS design tool is to design and implement a custom IP core module for the convolutional neural network in the image classification and recognition system. After the design of the custom IP core is completed, the IP core module and the ZYNQ module are mainly used to implement the path construction of the overall system. After verification, the upper computer program is used to call the control in Jupyter Notebook. Finally, the design of the driver and the host computer is completed, and the system is tested for functions and performance. The test shows that MNIST and CIFAR-10 data sets can be recognized and classified by the system, and the power consumption of the system is only 1.54 W. The image classification and recognition system has the advantages of versatility, energy-saving, etc., which can be widely used in edge computing environment.

Key words: convolutional neural network; software and hardware co-design; PYNQ; VIVADO; Jupyter Notebook

0 引 言

人工智能(artificial intelligence, AI)在计算机科学技术高速发展的今天逐渐成为各国的重点研究领域,其发展水平将很大程度决定一个国家的技术战略高

收稿日期:2020-11-21 **修回日期**:2021-03-23
基金项目:江苏省研究生科研创新计划(KYCX20_0803);南京邮电大学国自孵化项目(NY220013)
作者简介:陈 禹(1994-),男,硕士研究生,研究方向为智能信号处理;通讯作者:孙科学(1981-),男,博士,教授,硕导,研究方向为模式识别技术、嵌入式系统与智能信号处理。

度[1]。作为人工智能的一个重要研究领域,图像识别技术起源于 20 世纪 40 年代,由于当时基础理论与硬件性能不足未能得到快速发展。到了 90 年代,支持向量机和人工神经网络的结合促进了图像识别技术的发展,其中卷积神经网络算法的出现使得人脸识别、图像分类、特征提取等领域出现了众多显著的成果[2-5]。

卷积神经网络虽然性能卓越,但由于卷积过程伴随着庞大的计算量,因此对设备的计算能力有很高的要求。现在主流的硬件平台包括 CPU、GPU、FPGA 和 ASIC 芯片[6-7],其中 CPU 拥有很高的灵活性,但由于自身设计架构的局限性,难以支持并行运算从而导致处理效率不足;而 GPU 价格昂贵,功耗太高,无法应用于嵌入式移动终端[8];专用集成电路(application-specific integrated circuit,ASIC)可以实现计算力和功耗的平衡,但是其定制化的特点导致通用性差、成本高昂且可迁移性低[9];而现场可编程逻辑门阵列(field programmable gate array,FPGA)配置了众多逻辑单元可用于深度学习算法的并行计算,其计算力强、功耗低、开发周期短,在面向嵌入式移动终端的应用场景上可以发挥其独有的尺寸、性能和功耗上的综合优势[10-12],且 FPGA 由于可重新编程的特点,可以为不同的应用场景提供相应的支持[13-15]。

因此,无论是为了高效地完成人工智能技术的应用,还是为了部署灵活低功耗的嵌入式移动终端设备,基于 FPGA 平台的系统都可以在保障计算性能的同时控制能耗。文中设计的 PYNQ 图像分类识别系统一方面可以大幅降低能耗,另一方面可以通过软件操作来实现对不同数据集的分类识别,有效提高系统灵活性,降低部署成本。

1 系统总体设计

首先,对整体系统进行分析,通过软硬件协同设计思路明确功能模块,完成任务划分。在处理系统(process system,PS)部分,通过 Jupyter Notebook 平台基于 Python 实现上位机程序二进制特征参数的读取以及对硬件的控制。在可编程逻辑(programmable logic,PL)部分,实现卷积神经网络的模块设计和系统通路搭建。

PYNQ 图像分类识别系统能够处理的图像尺寸应在 PYNQ 有限的片上资源限制下实现灵活处理的设计要求,文中将通过 MNIST 数据集与 CIFAR-10 数据集验证系统的通用性,通过 Jupyter Notebook 开发环境进行系统测试。由于移动端嵌入式平台对系统的可移植性以及整体功耗有较高的要求,因此本系统需要对这两个方面进行针对性设计,其中对于系统功耗的设计要求不高于 5 W 以满足低功耗设计要求。图像分类识别系统的构成示意图如图 1 所示。

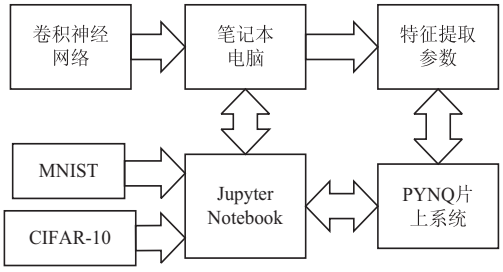

图 1 图像分类识别系统构成示意图

2 卷积神经网络的研究与设计

2.1 卷积神经网络的基本结构

卷积神经网络有前向传播和反向传播两种传播方式,网络结构主要由卷积层(convolutional layer)、池化层(pooling layer)、激活函数层(activation function layer)、Dropout 层、批规范化层(batch normalization layer)和全连接层(fully connected)构成[16-17]。其中,卷积层和全连接层主要完成特征提取,而池化层和 Dropout 的主要目的是防止过拟合现象的出现。

2.2 卷积神经网络的设计流程

图像分类识别系统中需要在电脑端完成卷积神经网络模型的设计与训练,最终处理得到特征参数二进制文件。

(1)针对 MNIST 数据集的网络设计。

MNIST 数据集的输入图片为 28 * 28 分辨率的灰度图,网络模型中第一个卷积层应用 ReLU 激活函数,然后是一个最大池化层。第二个卷积层通过 ReLU 激活函数和最大池化层对其进行处理,最后使用两个全连接层处理。MNIST 卷积神经网络的结构模型如图 2 所示。

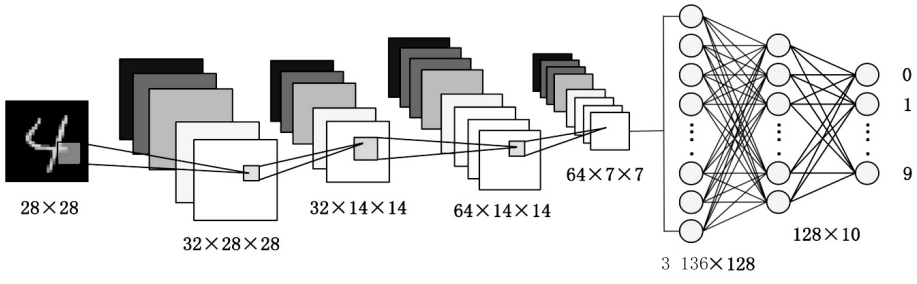

图 2 MNIST 数据集模型架构

建模过程中,采用自适应估计算法(adaptive moment estimation,Adam)来调节连接权值和偏置,最终测试识别准确率为 99.06%。

(2)针对 CIFAR-10 数据集的网络设计。

CIFAR-10 数据集的输入图片为 32 * 32 分辨率的彩色图,对模型的第一个卷积层添加 ReLU 激活函数,然后是最大池化层。第二个卷积层同样通过 ReLU 激活函数最大池化层进行处理。在两个卷积层和池化层之后是全连接层,最后是输出层。CIFAR-10 卷积神经网络的结构模型如图 3 所示。

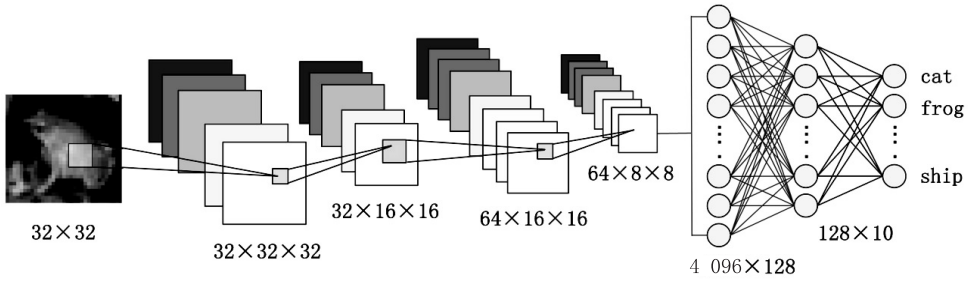

图 3　CIFAR-10 数据集模型架构

由于 CIFAR-10 数据集的计算量大幅提高,因此使用深度更高的网络进行学习训练就对硬件性能有较高的要求。由于文中所用电脑的硬件性能有限并且 CIFAR-10 数据集主要用于验证系统的通用性特点,因此所设计的网络为轻量级结构。最终 CIFAR-10 数据集测试准确率为 72.93%。

(3)特征参数的提取与格式转换。

在搭建神经网络模型后需要对网络中卷积层与全连接层的权重参数与偏执参数进行提取与格式转换。在神经网络程序中定义参数提取函数,根据维度的不同定义子函数,在提取函数中对参数进行维度确认并通过子函数完成参数提取,其中数据的先后读取顺序必须保持顺序相同,针对不同维度的数据,由维度从高到低存储,以便在 PYNQ 环境中运行正常。代码顶部先通过 sess = tf. InteractiveSession( ) 声明会话,对提取到的特征参数通过 tofile 函数定义保存路径及格式。完成特征参数的提取及格式转换后,得到二进制格式的特征参数文件。

3　基于 PYNQ 的图像分类识别系统实现

3.1　IP 核模块设计

(1)卷积层 IP 核模块参数配置。

在卷积运算中卷积核与输入数据进行相乘再累加的操作,卷积核通过给定的步长参数进行位移。因此在 HLS 的卷积层函数编程中首先需要定义输入输出参数以及卷积核及其步长等参数,输入输出特征参数数据类型为多维矩阵,使用指针地址形式传入数据的首地址。卷积层参数设置如表 1 所示。

(2)卷积层 IP 核模块设计。

在卷积层 IP 核模块设计中,设计流程严格遵循 HLS 开发工具的相关要求,首先添加头文件和 C 代码文件。在 C 代码文件中,需要首先设定函数 Conv 为 IP 核的顶层函数,在此函数中的各个特征参数定义在表 1 中已给出。在定义参数后需要对 IP 核模块的各个参数接口进行设计配置。

表 1　卷积层参数设置

参数含义	参数定义	参数字长
输入图像长度	l_in	ap_uint<8>
输入图像宽度	w_in	ap_uint<8>
输入通道数	c_in	ap_uint<9>
输出通道数	c_out	ap_uint<9>
卷积核长度	kernel_x	ap_uint<5>
卷积核宽度	kernel_y	ap_uint<5>
卷积核横向步长	kx_stride	ap_uint<3>
卷积核纵向步长	Ky_stride	ap_uint<3>
横向 padding	padding_x	ap_uint<4>
纵向 padding	padding_y	ap_uint<4>

然后选择卷积层 C 代码程序作为实现目标,设置仿真综合的时钟周期,对工程进行 C 仿真测试,编译并执行程序,验证无误后进行算法综合与 C/RTL 协同仿真,综合结束后生成性能报告。最后通过 HLS 执行 Export>RTL 将 IP 核模块以 RTL 形式生成即可看到 IP 核模块的 BLOCK,如图 4 所示。

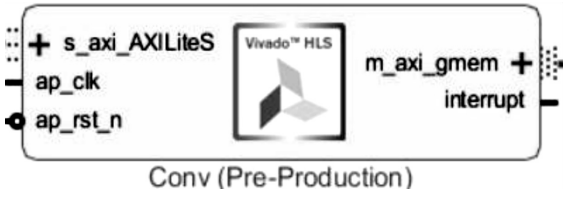

图 4　卷积层 IP 核通用电路

(3)最大池化层 IP 核模块设计。

与设计卷积层函数相同,最大池化层 IP 核模块设计需要在 HLS 中定义输入输出参数以及卷积核及其步长等参数,输入输出特征参数数据类型为多维矩阵,同样使用指针地址形式传入数据的首地址。完成最大

池化层 IP 核模块的设计后,可以将设计好的 IP 核加载到 VIVADO 软件中进行下一步的硬件通路设计。

3.2 整体通路配置

(1)ZYNQ 芯片参数配置。

在 VIVADO 中导入卷积层 IP 核模块以及最大池化层 IP 核模块后,还需要配置 ZYNQ 芯片完成通路搭建。操作步骤如下:①对片上资源 I/O 设备接口进行初始化;②对 PS 端与 PL 端的接口进行配置;③设定时钟规范,在时钟配置上,本课题的 PS 部分时钟频率为 650 MHz,PL 部分时钟频率配置为 100 MHz。

(2)系统 I/O 口配置。

文中设计的系统需要 USB、SD 接口、URAT 等 I/O 设备,在 VIVADO 中对其进行接口配置,在 ZYNQ 芯片的 I/O 外设选项,可以直观地看到 ZYNQ 可以提供的外设接口图形化的配置界面。

3.3 驱动程序与上位机程序设计

(1)卷积层与最大池化层驱动程序设计。

PYNQ 开发平台的一个显著特点是其 PS 部分搭载的操作系统支持开发者直接使用 Python 编程语言对硬件模块进行控制操作,Xilinx 官方提供的函数库文件可以方便开发者通过 Python 开发硬件。

由于文中设计的卷积层 IP 核模块是面向片上存储器的物理地址并且需要在操作系统中通过 Python 脚本程序对硬件进行操控,因此数组的虚拟地址与物理地址就可能会产生差异,所以在新定义数组时必须保证分配的是连续内存物理地址。在 Xilinx 官方提供的库函数中,Xlnk 可以用于分配连续物理地址以及提供物理指针,而 Xlnk 可以使用 Python 的 Numpy 第三方函数库来分配数组,但在对各个特征参数写入物理地址前,需要在驱动中对各个特征参数进行初始化复制。

在完成特征参数的初始化后就需要通过 PYNQ 的 pynq. Overlay 模块将数据写入至待写入的物理内存地址中。通过 Overlay 中的 download( ) 方法将 VIVADO 中生成的比特流文件下载至 PL 部分,wirte(offset,value) 函数将要写入的数据 value(int 数据类型或字节数据类型)写入 offset 地址。由于在卷积层 IP 核模块中,输入输出特征、权值和偏执通过指针传入首地址,因此在驱动程序中需要通过 physical_address 函数传入分配的物理地址。最大池化层 IP 核模块的驱动程序设计方法与卷积层 IP 核模块的设计方法相同,故不再赘述。

(2)上位机程序设计。

为了能够在 Liunx 操作系统和 Jupyter Notebook 中控制系统正常工作,需要通过 Python 语言设计编写上位机控制程序,上位机程序包括主函数和系统的初始化。

主函数中,首先读取灰度图/RGB 图,接着调用驱动函数,然后定义参考值 max 并赋值,最后将输出分类概率值与 max 对比,将最高概率值赋值为 max 并作为结果输出。系统初始化的程序结构如图 5 所示。

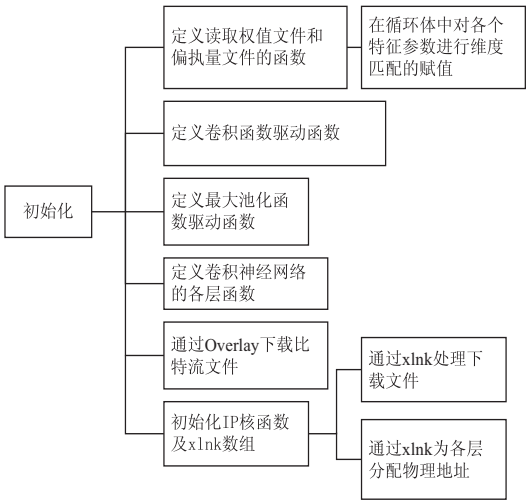

图 5 初始化程序结构

4 系统测试

图像分类识别系统搭载在 PYNQ-Z2 开发板上,在进行测试验证前需要对开发板硬件环境以及软件环境正确配置,将 OVERLAY 烧写至开发板中并利用 Python 编程语言实现控制操作。采用 MNIST 及 CIFAR-10 数据集对系统进行性能测试,验证文中设计并与桌面 CPU 和 GPU 性能及功耗进行对比分析。测试结果如表 2 所示。

表 2 平台横向参数对比

实验平台	MNIST 准确率/%	CIFAR-10 准确率/%	功耗/W	成本/元
Xeon E5-2692	99	80	135	15 000
Tesla K20	99	80	225	24 000
PYNQ-Z2	99.06	72.93	1.54	1 000

由表 2 可以看出,在 MNIST 数据集的识别率方面,PYNQ 图像分类识别系统与 E5-2692 和 K20 平台几乎相同,在 CIFAR-10 数据集的识别率上则仍有较大差距,但在功耗和成本方面本课题所设计的系统具备较明显的优势。与各平台的实现方式相比,PYNQ 图像分类识别系统不需要针对不同的数据集对系统进行重构,通过上位机程序读取不同测试集的特征参数,就可以实现不同测试集的正常分类识别。测试表明系统的分类识别功能与通用性符合设计要求。

5 结束语

设计了一种基于 PYNQ 开发平台的图像分类识

别系统,该系统以软硬件协同设计理论作为指导思想,在基于 ZYNQ 架构的 PYNQ-Z2 平台上设计了 PS 部分的驱动和上位机程序,在 PL 部分设计了自定义 IP 核模块并完成系统搭建。最后在 Jupyter Notebook 平台通过两个数据集进行功能与性能测试,通过测试可知,该系统可以在低功耗的前提下实现对不同数据集的分类识别。

为了进一步提高系统的性能,下一步可以尝试更多的算法实现,例如针对卷积神经网络模型的搭建,可以使用二值化卷积神经网络对数据进行压缩以降低存储消耗,此外还可以尝试 YOLO 算法作为识别方法。

参考文献:

[1] SOKOLOV I A. Theory and practice of application of artificial intelligence methods[J]. Herald of the Russian Academy of Sciences,2019,89(2):115-119.

[2] LECUN Y,BOTTOU L,BENGIO Y,et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE,1998,86(11):2278-2324.

[3] WU J,YIN X,XIAO H. Seeing permeability from images: fast prediction with convolutional neural networks[J]. Science Bulletin,2018,63(18):1215-1222.

[4] 黄凯奇,任伟强,谭铁牛. 图像物体分类与检测算法综述[J]. 计算机学报,2014,37(6):1225-1240.

[5] SALAKHUTDINOV R, HINTON G. An efficient learning procedure for deep Boltzmann machines[J]. Neural Computation,2012,24(8):1967-2006.

[6] BADAR M,HARIS M, FATIMA A. Application of deep learning for retinal image analysis: a review[J]. Computer Science Review,2020,35:100203.

[7] PHAN H,HERTEL L,MAASS M,et al. Robust audio event recognition with 1-max pooling convolutional neural networks[J]. arXiv:1604.06338,2016.

[8] VEDALDI A,JIA Y,SHELHAMER E,et al. Convolutional architecture for fast feature embedding[J]. arXiv: 1408.5093,2014.

[9] 孟李林. FPGA 和 ASIC 设计特点及应用探讨[J]. 半导体技术,2006,31(7):526-529.

[10] 杨雨诺,张国林,孙科学,等. 基于深度学习网络的心音智能分析平台构建[J]. 计算机技术与发展,2019,29(7):130-134.

[11] 卢　冶,陈　瑶,李　涛,等. 面向边缘计算的嵌入式 FPGA 卷积神经网络构建方法[J]. 计算机研究与发展,2018,55(3):551-562.

[12] 刘龑铭,孙科学,王淑媛,等. 基于 Nios Ⅱ 的 RFID 物流管理系统设计与实现[J]. 计算机技术与发展,2016,26(10):142-145.

[13] STALLKAMP J,SCHLIPSING M,SALMEN J,et al. Man vs. computer:benchmarking machine learning algorithms for traffic sign recognition[J]. Neural Networks,2012,32:323-332.

[14] FERN N,SAN I,KOÇ C K,et al. Hiding hardware trojan communication channels in partially specified soc bus functionality[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,2016,36(9):1435-1444.

[15] 李学龙,史建华,董永生,等. 场景图像分类技术综述[J]. 中国科学:信息科学,2015,45(7):827-848.

[16] 刘　双. 基于卷积神经网络的图像分类算法研究[D]. 昆明:云南财经大学,2020.

[17] 王　统. 深度学习中卷积神经网络的结构及相关算法[J]. 信息与电脑,2020,32(8):41-43.

(上接第 49 页)

计算模型[J]. 计算机研究与发展,2017,54(5):907-924.

[8] 黄倩怡,李志洋,谢文涛,等. 智能家居中的边缘计算[J]. 计算机研究与发展,2020,57(9):1800-1809.

[9] 施巍松,张星洲,王一帆,等. 边缘计算:现状与展望[J]. 计算机研究与发展,2019,56(1):69-89.

[10] GLOROT X, BENGIO Y. Understanding the difficulty of training deep feedforward neural networks[J]. Journal of Machine Learning Research,2010,9(1):249-256.

[11] KRIZHEVSKY A, SUTSKEVER I, HINTON G. Imagenet classification with deep convolutional neural networks[C]// Advances in neural information processing systems. New York,USA:ACM,2012:1097-1105.

[12] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large scale image recognition[J]. arXiv:1409.1556,2014.

[13] KIM Y D,PARK E,YOO S,et al. Compression of deep convolutional neural networks for fast and low power mobile applications[J]. arXiv:1511.06530,2015.

[14] ZHANG K,YANG Z,BAŞAR T. Multi-agent reinforcement learning:a selective overview of theories and algorithms[J]. arXiv:1911.10635,2019.

[15] TOKUI S, OKUTA R, AKIBA T, et al. Chainer: a deep learning framework for accelerating the research cycle[C]// Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. New York, USA:ACM,2019:2002-2011.