

基于 FPGA 的病虫害图像识别系统研究

曾 沛

辽宁工程技术大学理学院, 辽宁 阜新 123000

摘 要: 针对农业生产中病虫害识别实时性要求高、传统深度学习模型在嵌入式设备部署困难的问题, 本文提出一种基于 FPGA 的改进 AlexNet 图像识别系统。通过引入批量归一化、精简全连接层结构, 优化模型以适应嵌入式环境; 设计专用卷积加速器, 结合循环分块与并行处理技术, 在 Zynq-7010 平台上实现硬件加速。实验采用自建苹果病虫害数据集, 通过数据增强与模型训练, 系统在 FPGA 部署后准确率与软件版本基本一致 (训练集 89.91%, 验证集 81.29%), 同时推理时间从 3211ms 缩短至 2537ms, 显著提升识别效率。本研究为轻量化 CNN 在农业边缘计算中的应用提供了可行的技术路径。

关键词: FPGA; AlexNet; 卷积加速器; 流水线优化

0. 绪论

病虫害识别对农业发展极为关键, 它能够控制病害的流传, 削减作物损失, 还能优化农作物的产量和品质, 传统的办法依靠人工设计特征, 既繁杂又依靠专业知识, 这约束了它的应用范畴, 卷积神经网络 (CNN) 凭借其出色的图像分类性能, 被普遍应用到病虫害识别当中, 从而明显改善识别精确度, 以最近几年为例, 很多改良过的 CNN 模型在识别水稻, 玉米和小麦病虫害时, 得到了超出 97% 的识别精度。

但是由于病虫害识别经常要部署在像无人机和智能相机这类嵌入式设备上而这些设备上计算资源较少所以像 AlexNet 和 VGG 这样需要消耗大量计算量的深层网络就不适合用在实时处理要求高的嵌入式设备上。为此本文提出了一种 FPGA 改进的 AlexNet 加速方法。利用数据量化、模型稀疏化来减少计算量, 设计专用硬件的卷积加速器, 采用并行和流水线加速来提高卷积处理速度。在 FPGA 上实现高效的推理。识别速度快很多, 能效也比较高, 给嵌入式环境中病虫害识别实时性提供了新的可能

1. AlexNet 网络的搭建

1.1 AlexNet 网络

近些年来, 深度学习模仿人脑神经网络创建多层模型, 达成对繁杂数据的识别, 传统神经网络在处理高复杂度任务的时候, 计算量大, 耗费的资源也比较多, 为解决这个难题, 卷积神经网络 (CNN) 采用卷积, 池化以及权重共享手段, 在有效提取特征的同时削减计算量, 而且优化了泛化能力。

CNN 基本流程包括: 卷积层提取特征, 池化层压缩特征并降维, 全连接层完成分类, 训练中采用梯度下降法优化权重。

AlexNet 是 CNN 发展中的重要里程碑, 由 Hinton 与 Alex Krizhevsky 提出, 在 2012 年 ImageNet 竞赛中以错误率 16.4% 夺冠。该网络共 11 层, 包括 5 个卷积层、3 个池化层和 3 个全连接层, 卷积核尺寸逐层递减, 并利用填充维持特征图尺寸。其引入 ReLU 激活函数, 显著提升训练速度, 缓解梯度消失, 增强网络稀疏性, 从而提升识别精度并减轻过拟合。

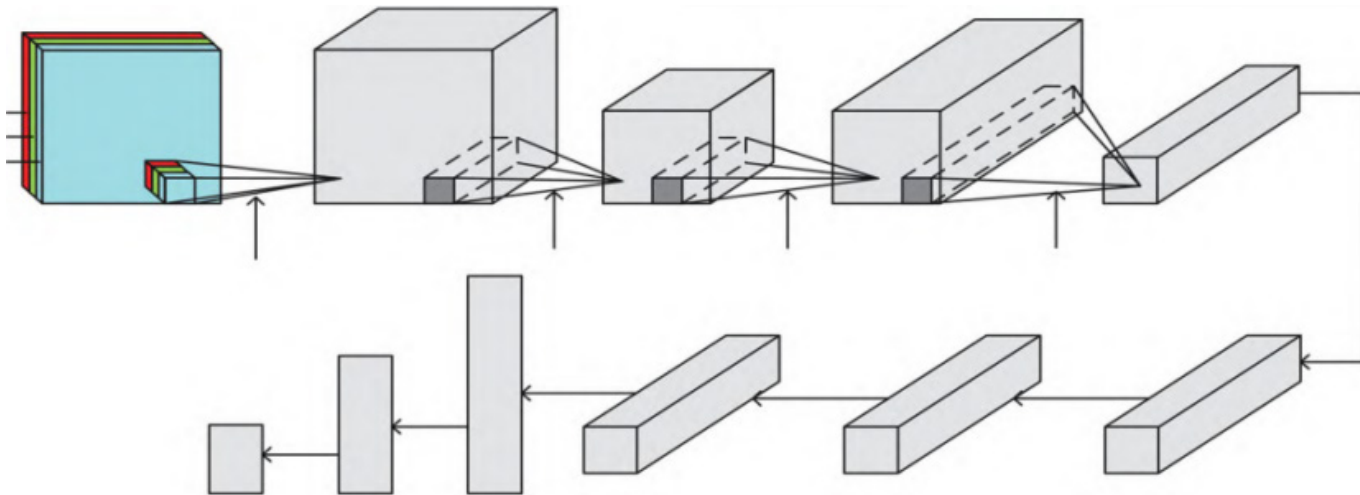


图 1 AlexNet 模型实现

1.2 AlexNet 模型实现

1.2.1 数据集

本文采用包含 664 张图像的苹果病虫害数据集, 涵盖苹果疮痂病、黑腐病、松柏锈病和健康苹果四个类别。为适配模型输入并提升泛化能力, 所有图像统一调整为 277×277 像素, 并通过过曝、平移、旋转等数据增强操作增加样本多样性, 有效提高了模型的鲁棒性和识别性能。

1.2.2 模型的搭建

为满足病虫害识别的嵌入式部署需求, 本文对 AlexNet 网络进行三方面改进: 增加批量归一化层以加速收敛并抑制过拟合; 将输出层神经元调整为 4 个以适配四分类任务; 缩减全连接层规模以降低资源占用。

改进后网络以 227×227 彩色图像为输入,

经五个卷积层与三个池化层提取特征。卷积核尺寸依次为 11×11 (96 个)、 5×5 (256 个) 和 3×3 (384/384/256 个), 每层卷积后均进行批量归一化。特征图经扁平化后, 由两个 1024 神经元的全连接层整合, 最终通过 4 神经元输出层完成分类。该结构在保证精度的同时显著提升了在 FPGA 上的部署效率。

1.2.3 模型的训练

模型的搭建在 jupyter 中实现, 在 TensorFlow 环境中, 利用 Keras 框架来搭建。将数据的 80% 作为训练集, 其余的 20% 作为验证集。通过 keras 框架下的函数搭建好改进后的 AlexNet 模型, 每次训练喂入 128 个数据, 经循环训练 60 后, 训练集的准确率为 93%, 验证集准确率为 82%, 具体训练结果如图 2 所示。

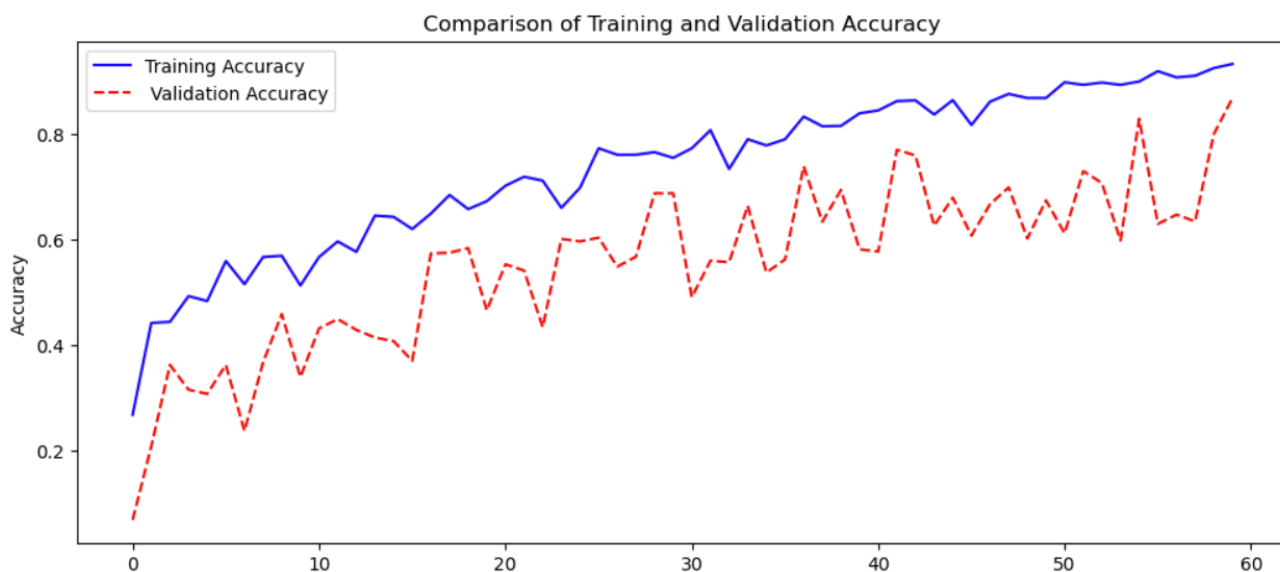


图 2 模型训练结果

将验证集效果最好的一次训练结果的权重进行提取, 并将其数据类型由浮点型转变为二进制数据并保存到文件中, 以便于 fpga 的读取。

2 FPGA 卷积加速器设计

2.1 卷积加速器

卷积是机器学习中的基础运算, 尤其在卷积神经网络中占据核心地位, 但其多重循环嵌套的特性导致计算复杂度高、耗时较长。为在 FPGA 等资源受限的平台上实现高效推理, 本文通过硬件加速方式对卷积运算进行优化。

借助 Xilinx 高层次综合工具, 将 C++ 代码转换为硬件描述语言, 设计出专用的卷积加速器。通过流水线处理、并行优化及循环分块等技术, 显著提升了运算效率。最终, 在 Zynq-7010 开发板的 PYNQ-Z1 环境中完成部署与验证, 有效实现了卷积层的硬件加速。

2.2 卷积加速器的实现

2.2.1 使用 FPGA 加速卷积运算

本文选用 FPGA 实现对卷积运算的加速, FPGA 的中文全称为现场可编程门阵列, 其可以根据用户的需要被重新编程为不同的硬件逻辑电路, 在进行卷积加速器设计时, 我们可以根据我们模型卷积核数量, 通道数量, 输入特征图大小等参数以及实验所需配置去设计 FPGA 的硬件电路, 并将指定循环计算在硬件电路展开, 从而实现卷积加速器设计上的灵活性与并行性。同时, 在执行相同的计算任务时, fpga 往往通常能以更低的功耗去完成。这些优点使得 FPGA 其更加适合部署到移动或者远端应用上。

2.2.2 搭建一个最简单的卷积加速器

在卷积神经网络中, 卷积层的实现如图所示

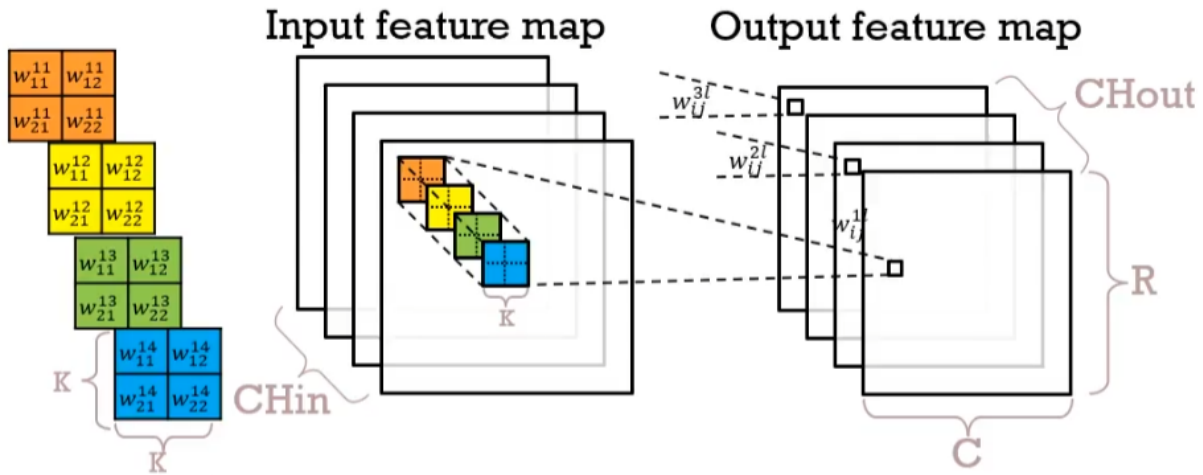


图 3 卷积层结构图

整个卷积层的输入是 $CHin$ 张输入特征图，输出是 $CHout$ 张输出特征图，每张输出特征图的大小为 $R \times C$ ^[1]。由于每一个“输入-输出”特征图对都有一个特定的卷积核用于卷积计算，所以总共有 $CHout \times CHin$ 个卷积核，每个卷积核的大小为 $K \times K$ 。在进行卷积计算的过程中，

卷积的运算公式如下：

$$Out[ch\ o][r][c] = \sum_{ch=0}^{CHin-1} \sum_{r=0}^{K-1} \sum_{c=0}^{K-1} In[cho][r+kr][c+kc] \times W[cho][chi][kr][kc]$$

2.2.3 循环展开 (Loop Unrolling)

循环展开 (Loop Unrolling) 是一种常见的编译器优化技术，其被广泛的应用在硬件设计当中，FPGA 可以通过配置其内部逻辑块和连接来实现不同的数字电路，并且其具有高度并行性，循环展开在 FPGA 设计中尤其有用。

在神经网络加速器中，最基本的计算单元被称为 PE (Processing Element)，每一个 PE 都需要并行读入 $CHin$ 个输入和 $CHin$ 个权重参数，同时处理 $CHin$ 个乘法，并将结果累积到一个 $CHout$ 通道上，因此我们需要 $CHout$ 个相同的 PE 同时运算，从而实现了不同输出通道计算的并行处理。因此，对应到代码上，需要将前两个循环进行循环展开。

2.2.4 循环分块及卷积核参数的确定

AlexNet 的五层卷积层通道数分别为 96、256、384、384 和 256。若将输入输出通道全展开计算，FPGA 资源将严重不足。为此，我们采用循环分块技术，每个时钟周期仅并行处理一小块计算 (PE 单元)，以在硬件资源与运行效率间取得平衡。

实验基于 Zynq7010 开发板，结合其硬件资

每个卷积核滑过各自的输入特征图^[2]，并使用当前滑过的窗口中的输入特征与卷积核内的权重完成卷积计算 (对应位置相乘，所有乘积累加)，卷积的结果会累加到对应位置的输出特征上。

源与卷积层特征，将每周周期处理的卷积块大小设定为 16。对该循环分块设计进行综合，结果验证了该方案在有限资源下的可行性。

表 1: 算法资源占用

Name	BRAM_18K	DSP48E	FF	LUT
DSP	—	—	—	—
Expression	—	—	0	168
FIFO	—	—	—	—

可以看到在不超过 fpga 板载资源的同时，尽可能高的提高了该模块的计算并行度从而即实现了硬件的固定，也可以根据不通的配置实现灵活调度。

2.3 实验与结果

将在第二章中训练完成的模型部署到 pynq 中，将设计的卷积加速器，通过 overlay() 函数挂载到 cpu 上，并重写 AlexNet 模型中的卷积层，将卷积层中的卷积计算传递到卷积加速器中进行运算。

将之前训练好的权重加载到 AlexNet 模型当中，对数据集的训练集以及验证集进行识别，其中的部分结果如下图所示例：



图 5 部分结果

得到的实验结果准确率与之前在 python 上面训练的结果准确率进行对比, 如表 2 和表 3 所示:

表 2: 识别准确率对比		
平台	训练集	验证集
python	90.29%	82.12%
pynq	89.91%	81.29%

在 pynq 的 ARM 上通过 cpu 进行预测, 并统计其计算所花费时间, 与挂载 FPGA 卷积加速器计算所花费时间进行对比, 如表 2 所示:

表 3: 运算速度对比 单位: ms	
平台	AlexNex 预测
ARM	3211
ARM+FPGA	2537

可以看出, zynq 进行硬件加速后, 与 python 软件版本相比准确率相差不大, 但运算速度上,

纯 ARM 的计算速度要远远落后于通过挂载着卷积加速器的 ARM 核。由此可以得出本文所设计的卷积加速器实现了既定目标, 加速了模型的计算速度, 降低了功耗。

3. 结论

本文成功设计并实现了一套基于 FPGA 的病虫害图像识别系统, 通过改进 AlexNet 结构、设计专用卷积加速器, 并结合循环分块与并行优化策略, 在 Zynq-7010 开发板上完成了硬件部署与验证。实验结果表明, 该系统在保持较高识别准确率的同时, 显著提升了推理速度, 验证了 FPGA 加速在嵌入式农业视觉任务中的有效性与实用性。未来将进一步优化模型结构与硬件资源调度, 探索动态可重构加速架构, 以拓展系统在复杂农田环境中的适用性与实时性能。

参考文献:

- [1] 刘拥民, 胡魁, 聂佳伟, 等. 基于 MSDB-ResNet 的水稻病虫害识别 [J]. 华南农业大学学报, 2023, 44(06): 978-985.
- [2] 武魁, 高丙朋. 改进 EfficientNetV2 的玉米病虫害识别与分级研究 [J]. 现代电子技术, 2023, 46(14): 68-74.
- [3] 王汉生, 姚建斌. 基于 ResNet 和 ViT 双流网络的小麦病虫害识别 [J]. 农业技术与装备, 2024, (02): 18-21.
- [4] 刘颀, 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(01): 121-125.

作者简介: 曾沛 (2005.5—), 男, 汉族, 湖南省, 本科生, 研究方向: 模糊数学, 计算机视觉。

基金项目: 辽宁省“兴辽英才计划”项目资助 (编号: XLYC2211085); 辽宁省基本科研项目 (编号: LJ242510147005; 编号: JYTQN2023210)。