

基于 PYNQ 的轻量化实时目标检测技术

郭向楠¹

GUO Xiangnan

摘要

针对边缘端设备在资源受限条件下部署深度学习目标检测算法的需求,文章提出了一种基于 PYNQ (python productivity for zynq) 框架的轻量化实时目标检测技术,开展 YOLOv5n 目标检测算法研究及适应 FPGA 部署的模型优化,完成 PYNQ 环境下的软硬件协同设计,并在 ZCU104 开发板上完成了系统验证。结果表明,所提技术能够实现轻量化深度学习目标检测算法的快速部署与验证,具有良好的检测精度与实时性,能效比达到 10.5 帧/(s·W),在低成本实时目标检测领域具有较高的应用价值。

关键词

目标检测; PYNQ; YOLOv5n; 轻量化; 边缘端设备

doi: 10.3969/j.issn.1672-9528.2025.03.009

0 引言

近年来,人工智能技术在全球范围内飞速发展,已从科技前沿探索阶段,逐步迈入规模化应用阶段。计算机视觉作为人工智能领域的重要分支,在工业检测、医学分析、安防监控及智能驾驶等众多领域发挥着重要作用^[1]。目标检测是计算机视觉的核心任务之一,主要用于识别图像或视频中目标的类别及位置。传统目标检测算法主要通过人工特征提取结合经典机器学习分类器实现^[2]。借助深度学习技术,新型目标检测算法开始具备自动特征提取及训练能力,发展出两阶段检测(two stage detection)、单阶段检测(one stage detection)及无锚框检测(anchor free detection)等多种类型^[3-5]。深度学习目标检测算法相对于传统算法在复杂场景检测精度及新任务的适应能力上具有较大的优势,但由于其处理流程复杂,对硬件及数据的要求较高,在实践应用中需要综合考虑性能、资源及成本等多个因素,同时针对具体使用场景进行优化。本文提出一种基于 PYNQ (python productivity for zynq) 框架的轻量化实时目标检测技术,开展 YOLOv5n 目标检测算法研究、适应 FPGA 部署的模型优化及 PYNQ 环境下软硬件协同设计,实现深度学习目标检测算法在边缘端设备轻量化、高能效比部署。

1 总体方案

围绕 PYNQ 框架实现深度学习目标检测算法开展总体方案设计。PYNQ 是 XILINX 公司推出的一种开源框架,将 FPGA 的硬件可编程特性与 ARM 的软件灵活性相结合,便于嵌入式系统的敏捷开发。PYNQ 在硬件层(PL 端)将

FPGA 实现的逻辑功能封装成 IP 核形式,在软件层(PS 端)运行 Linux 系统,同时提供可视化交互式开发环境 Jupyter Notebook,使用 Python 调用应用程序编程接口(application programming interface, API)对硬件进行操作。其系统架构如图 1 所示。

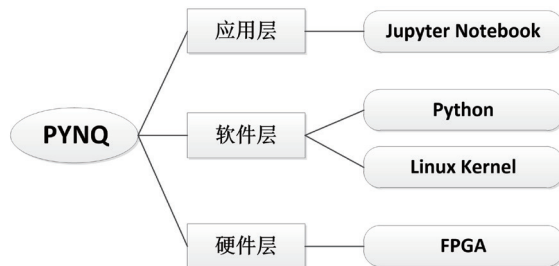


图 1 PYNQ 系统架构

硬件采用 PYNQ 支持的 XILINX ZCU104 开发板,利用四核 Arm Cortex-A53 及双核 Cortex-R5 处理器进行通用计算机逻辑控制,配合具备高密度逻辑资源和 DSP 模块的 UltraScale+ FPGA 进行神经网络加速,同时利用大容量高速 DDR4 及丰富的外设接口实现数据交互,能够满足深度学习目标检测算法的硬件需求。

算法选择上主要考虑边缘端设备的使用需求及部署难度。在各类算法中,以方向梯度直方图(histogram of oriented gradients, HOG)+支持向量机(support vector machine, SVM)组合为代表的传统算法对先验知识依赖高,检测精度及适用场景有限^[6]。以更快速局部卷积神经网络(Faster R-CNN)为代表的两阶段检测算法将目标检测拆分为候选区域生成及分类两个独立阶段,通过高质量筛选降低干扰,比较适合复杂背景及小目标检测^[7],但整体计算量较大,延迟偏高。以 CenterNet 为代表的无锚框检测算法利用关键点直接预测边界

1. 中国空空导弹研究院 河南洛阳 471000

框,降低后处理难度^[8],但相关算子对FPGA的支持性不好。相对而言,YOLO(you only look once)系列的单阶段检测算法无须预先生成候选区域,通过单次向前传播直接预测目标分类及位置,尤其适合应用于实时目标检测^[9]。自2016年诞生以来,YOLO算法不断更新迭代。YOLOv1算法将图像划分为众多网格,通过每个网格预测多个边界框及置信度,其特点是模型简单且易于训练,但处理小目标及复杂场景能力较弱。YOLOv2算法引入锚框机制,支持多尺度训练提升定位精度,但模型复杂性有所提高。YOLOv3算法开始使用Darknet-53主干网络,引入多尺度预测机制,在速度和精度上取得了较好的平衡。YOLOv4算法升级了CSPDarknet53主干网络,减少计算量的同时提升训练效果。YOLOv5算法基于PyTorch框架实现,开始向高速与轻量化发展,尤其适合移动设备及嵌入式系统部署^[10]。因此,本文采用YOLOv5系列中最轻量级YOLOv5n版本,参数量仅1.7MB,计算量约0.5BFLOPs,同时具有良好的检测精度,适合边缘端设备部署。

2 模型转换与优化

YOLOv5n模型主要网络结构如图2所示。其中主干网络Backbone通过卷积(Conv)、跨阶段残差(C3中包含)及空间金字塔快速池化(SPPF)操作提取多尺度特征,颈部网络Neck通过上采样(Upsample)、通道连接(Concat)及跨阶段残差(C3中包含)进行多尺度特征图融合,检测头Head最终生成边界框、类别概率及置信度等检测结果。

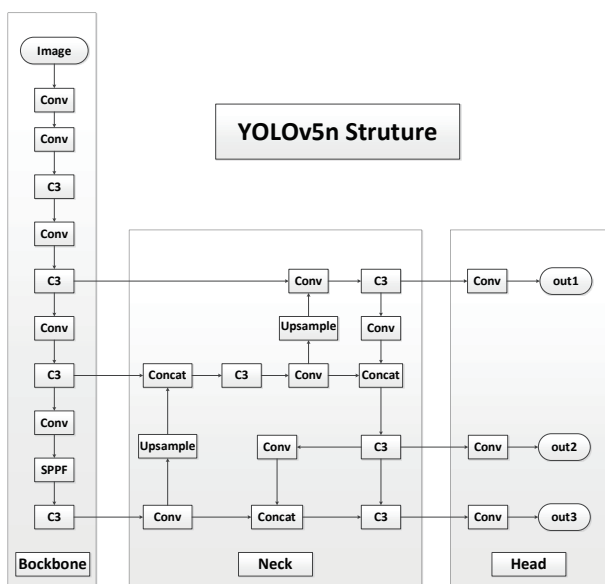


图2 YOLOv5n 网络结构

从网络结构中可以看出,由主干网络及颈部网络完成的特征提取及融合部分存在大量卷积及乘加操作,适合FPGA通过并行处理进行加速,检测头部分较为复杂的特征检测可

由处理器协助完成。

由于YOLOv5n官方开源预训练模型为PyTorch格式,为便于后续FPGA编译及PYNQ环境运行,首先需要将模型向Onnx格式转换。由于Onnx标准算子库不支持SiLU(sigmoid linear unit)激活函数,将模型中全部SiLU函数用Sigmoid函数加Mul乘法器进行替代。卷积层后的BatchNorm2d节点用于加速训练收敛并提升模型泛化能力,用于推理时,将其与之前的卷积层融合为带偏置的卷积,可以减少模型参数量,降低推理时的内存占用,同时不影响模型精度。颈部网络中,用于放大特征图的上采样部分在Onnx中映射为Resize节点,同时将C3结构中的残差连接映射为Add节点。转换前后模型局部对比如图3所示。

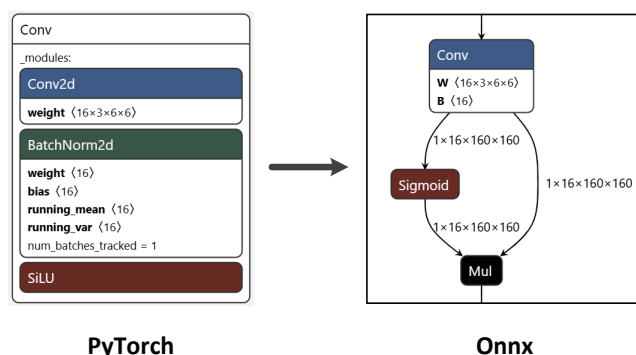


图3 由PyTorch到Onnx模型转换

转换完成后,对模型进行下一步的优化处理。由于本文仅使用FPGA实现主干网络及颈部网络的特征提取部分,检测部分由处理器完成,因此需要将模型进行分割。通过模型可视化工具导入转换后Onnx格式的模型文件,可以查看具体网络结构及节点。找到网络中最后一个卷积节点model.24,可以此为特征提取与特征检测分界点将模型网络分割,如图4所示。

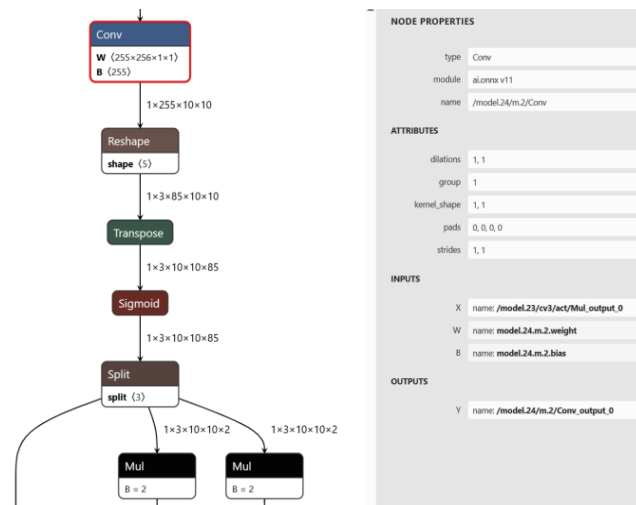


图4 特征提取与特征检测分界点

使用Graphsurgeon库对模型进行编辑,将model.24后的

模型保存为检测头模型,供处理器使用。同时将 model.24 及之前模型调整输入输出通道的位宽与 FPGA 设计匹配,生成新的特征提取模型。

3 PYNQ 环境下的软硬件协同设计

为了在 FPGA 上实现特征提取模型的功能,需要将 Onnx 模型转换为 FPGA 硬件描述文件。首先使用 Parser 库导入特征提取模型进行解析,同时导入 ZCU104 开发板的硬件配置 toml 文件,获取 FPGA 硬件资源信息。为确保模型能在 FPGA 上正确部署并提高运行效率,需要对关键参数进一步优化,如重新设定输入输出层的位宽及精度,调整不同卷积层的并行计算单元数量以优化数据流,根据卷积层参数文件大小分配不同类型的存储资源,导出模型硬件描述文件、参数文件及查找表,最终通过 FPGA 工具链生成 BIT 流文件。

在 ZCU104 开发板上配置好 PYNQ 环境,通过网线与上位机连接,使用 Jupyter Notebook 浏览器工具将所需文件导入开发板中。通过板上 USB 接口连接外置摄像头,获取实时图像输入。PYNQ 环境搭建如图 5 所示。



图 5 基于 ZCU104 开发板搭建 PYNQ 环境

PYNQ 的工作流程主要分为初始化配置、图像输入及预处理、特征提取、图像后处理及可视化等。开发板上电后,从 SD 卡加载预编译的 PYNQ 镜像文件,进入初始化配置阶段,完成相关依赖库调用、文件路径设置、配置参数读取及模型初始化等操作。通过硬件驱动模块控制 FPGA 预加载 Bit 流及参数文件,激活 FPGA 特征提取功能。图像输入及预处理阶段,调用 USB 摄像头进行实时图像采集,同时将图像进行尺寸缩放、通道转换及归一化,确保 FPGA 输入接口的适配。测试时也可修改配置文件,使用存储的图像作为输入。特征提取阶段,通过 FPGA 接口程序调配硬件资源,管理数据传输通道,使图像在 FPGA 中完成特征提取功能,并将结果输入处理器运行的检测头模型进一步推理。图像后处理阶段,对模型推理结果进行非极大值抑制(non maximum suppression, NMS),采用置信度阈值过滤低概率的预测框,采用交并比阈值控制重叠框的抑制程度,最终输出预测结果。为便于查看目标检测结果,通过可视化处理在原始输入图像

上绘制最终预测的类别名称及检测框,同时显示图像处理的总延时。

4 系统验证

使用 COCO128 标准数据集对本文方法与 YOLOv5n 完整 Onnx 模型直接推理方案的检测精度进行对比,结果如表 1 所示。

表 1 检测精度对比

	Precision	Recall	mAP@0.5	mAP@0.5:0.95
FPGA	0.591 4	0.344 9	0.372 8	0.223 6
Onnx	0.572 6	0.443 5	0.469 1	0.311 2

其中,查准率 Precision 代表预测为真样本的正确比例,查全率 Recall 代表实际为真的样本被正确预测的比例,mAP 代表不同 IoU 阈值下全部类别预测的平均精度。结果表明,FPGA 正确实施了特征提取功能,在查准率上略微提高,在查全率及平均精度上有所降低,与模型转换后的精度下降相关,后续可通过训练及参数调整进一步提高精度。

对单幅图片进行目标检测测试的结果如图 6 所示。结果表明,本文方法对复杂场景重合类目标的检测精度良好,单幅图像推理延迟约 61 ms,相比 Onnx 模型直接推理延迟降低一半以上,FPGA 并行处理的推理加速效果明显。



图 6 单幅图片目标检测效果

对 USB 摄像头采集图像进行实时目标检测结果如图 7 所示,能够正确识别常用物品的类别,检测框位置准确,处理延迟在 19 ms 左右,具有较好的实时性。文中使用的 ZCU104 开发板运行功耗约 5 W,能效比达到 10.5 帧/(s·W),能够满足边缘端设备的低功耗需求。

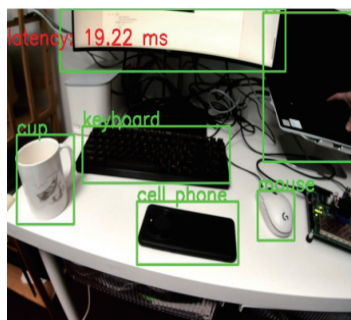


图 7 摄像头采集图像实时检测效果

(下转第 50 页)

统一规范的监控管理机制。下一步,API 网关将有序拓展应用场景,如一件事系统、电子证照共享平台、不动产登记等各类政务服务系统,将统一基于 API 网关提供接口调用服务,促进跨系统间安全有效的接口调用。

参考文献:

[1] 国务院关于进一步优化政务服务提升行政效能推动“高效办成一件事”的指导意见[J]. 中华人民共和国国务院公报, 2024(3): 15-22.

[2] 国务院办公厅关于加快推进电子证照扩大应用领域和全国互通互认的意见[J]. 中华人民共和国国务院公报, 2022(7): 31-35.

[3] 卜意磊. 微服务架构应用于市场监管应用支撑中心的研究[J]. 信息系统工程, 2020(12):72-73.

[4] 国务院关于加强数字政府建设的指导意见[J]. 中华人民共和国国务院公报, 2022(19):12-20.

[5] 方钟亮. 工程建设项目审批监管系统设计[J]. 现代信息技术, 2024, 8(22):151-155.

[6] 何平, 刘晓毅, 王进, 等. 容器云的 API 安全技术研究[J]. 保密科学技术, 2022(7):41-46.

(上接第 44 页)

5 结语

本文设计并实现了一种基于 PYNQ 环境的轻量化实时目标检测技术,开展了 YOLOv5n 目标检测算法研究及适应 FPGA 部署的模型优化,完成 PYNQ 环境下的软硬件协同设计,并在 ZCU104 开发板上完成了系统验证。结果表明,本方案能够实现轻量化深度学习目标检测算法的快速部署与验证,具有良好的检测精度与实时性,满足边缘端设备的低功耗需求,在低成本实时目标检测领域具有较高的应用价值。

参考文献:

[1] 卢宏涛, 张秦川. 深度卷积神经网络在计算机视觉中的应用研究综述[J]. 数据采集与处理, 2016,31(1):1-17.

[2] 李柯泉, 陈燕, 刘佳晨, 等. 基于深度学习的目标检测算法综述[J]. 计算机工程, 2022,48(7):1-12.

[3] 陈超, 齐峰. 卷积神经网络的发展及其在计算机视觉领域中的应用综述[J]. 计算机科学, 2019,46(3):63-73.

[4] 陈科圻, 朱志亮, 邓小明, 等. 多尺度目标检测的深度学习研究综述[J]. 软件学报, 2021,32(4):1201-1227.

[7] 普拉巴斯. 西里瓦德纳. API 安全进阶基于 OAuth2.0 框架: 原书第 2 版[M]. 李伟, 译. 北京: 机械工业出版社, 2022.

[8] 冯骐, 沈富可. 高校能力开放平台中的 API 网关设计与实现[J]. 中国教育信息化, 2021(3):61-66.

[9] 崔庆才. Python3 网络爬虫开发实战: 第 2 版[M]. 北京: 人民邮电出版社, 2021.

[10] 钱红兵, 李艳丽, 张蕊. WebCollector 和 ElasticSearch 在高校网站群敏感词检测中的应用研究[J]. 电子设计工程, 2019, 27(24): 11-14.

[11] 王志军, 姚文达. Web API 后端接口管理与应用[J]. 智能计算机与应用, 2024,14(5):247-251.

[12] 互联网政务应用安全管理规定[J]. 中华人民共和国公安部公报, 2024(3):7-11.

[13] 张虎, 张秋萍. 基于 KONG 的 API 集成系统的设计与实现[J]. 计算机技术与发展, 2019,29(8):102-106.

【作者简介】

方钟亮(1980—), 男, 江苏昆山人, 硕士, 高级工程师, 研究方向: 政务信息化建设。

(收稿日期: 2024-12-18)

[5] 刘俊明, 孟卫华. 基于深度学习的单阶段目标检测算法研究综述[J]. 航空兵器, 2020, 27(3): 44-53.

[6] 徐渊, 许晓亮, 李才年, 等. 结合 SVM 分类器与 HOG 特征提取的行人检测[J]. 计算机工程, 2016, 42(1): 56-60.

[7] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE transactions on pattern analysis and machine intelligence. 2017,39(6):1137-1149.

[8] 郑婷婷, 杨雪, 戴阳. 基于关键点的 Anchor Free 目标检测模型综述[J]. 计算机系统应用, 2020, 29(8): 1-8.

[9] 邵延华, 张铎, 楚红雨, 等. 基于深度学习的 YOLO 目标检测综述[J]. 电子与信息学报, 2022, 44(10): 3697-3708.

[10] 杨晓玲, 江伟欣, 袁浩然. 基于 YOLOv5 的交通标志识别检测[J]. 信息技术与信息化, 2021(4): 28-30.

【作者简介】

郭向楠(1988—), 男, 河南洛阳人, 硕士, 工程师, 研究方向: 嵌入式软件设计与测试。

(收稿日期: 2025-02-20)