基于 FPGA 硬件加速 CNN 目标识别

赖瀚昀　郑小海　王露曼
西京学院　陕西省西安市　710123

摘　要：为提高计算速度卷积神经网络（CNN）用于目标识别任务，本文研究了现场可编程门阵列（FPGA）在加速 CNN 目标识别中的应用。FPGA 以其硬件可重构性、低延迟和高效能效比，在模型的适应性方面展现出显著优势。本研究选用了 YOLOv3 作为 CNN 模型，通过 FPGA 实现目标识别的加速，通过软硬件协同工作流程，实现了深度学习模型的高效加速。本文介绍了基于 FPGA 的卷积运算设计，包括多通道卷积运算和 DSP48 完成卷积乘法的优化策略。实验结果表明，该系统表现出良好的目标检测效果，在资源消耗和功耗方面具有优势，低功耗特性更适合嵌入式系统。

关键词：现场可编程门阵列；目标识别；嵌入式系统；YOLOv3

引言：

随着人工智能技术的飞速发展及其在各领域的广泛应用，对硬件计算平台的需求在加强：既要具备强大的计算能力和高效的能效比，又要能够灵活适应不断涌现的各种 AI 算法和模型，特别是在卷积神经网络（CNN）这类深度学习模型用于目标识别的任务中。FPGA 凭借其硬件可重构性，能够在不更换硬件的前提下，快速实现不同硬件设计的加载和功能切换，这种“现场可编程”特性使其能够迅速响应不断变化的应用环境和技术进步。尤其是在 CNN 目标识别的实际应用中，推断阶段需要在保证高精度的同时，实现低延迟、高吞吐量以及对多种应用场景的广泛适应性。本研究将深入探讨 FPGA 在硬件加速 CNN 目标识别中的独特优势及其具体实现方案，力求为解决人工智能时代计算硬件面临的挑战提供新的视角和实践路径。本文使用经典的卷积神经网络模型 YOLOv3，并通道复用法，使加速功能可以在算力更低的 ZYNQ 开发板中运行，有力推动 CNN 目标识别技术在实际应用中的普及与深化。

1 深度学习原理与硬件搭建

1.1 YOLOv3 目标检测算法

本研究选用了 YOLOv3 进行部署。YOLOv3 是 YOLO 系列目标检测算法的第三个版本，它在 YOLOv2 的基础上做出了多项改进和优化，YOLOv3 使用新的特征提取网络 Darknet-53，提高了特征提取的能力，使 YOLOv3 在检测精度上有了显著的提升。并且引入了残差连接（Residual Connections），这有助于训练更深的网络模型，同时保持较高的处理速度。并且通过向网络的不同层添加预测层。这种策略使得 YOLOv3 能够更好地检测不同大小的物体。

1.2 FPGA 硬件部署

本次设计硬件为 Xilinx FPGA ZYNQ-7030 的开发板，ZYNQ-7030 芯片内嵌有两个 ARM Cortex-A9 处理器，它们可以运行 Linux 或 Android 操作系统，支持多线程和多任务处理。除了 ARM，ZYNQ-7030 还包含可编程逻辑单元，允许开发者根据特定应用需求设计和实现自定义硬件加速器。其结合了 ARM 的软件处理能力和 FPGA 的硬件加速能力。通过硬件加速可以显著提高特定任务的处理速度，特别是在需要并行处理或实时处理的场合。

2 加速方案部署和实现

2.1 ZYNQ 加速方案

在本方案中，核心是将 PS 端作为整个深度学习模型的调度中心，负责数据预处理、模型加载以及最终结果的解析。由于 PL 端在运算和存储能力上存在一定的局限性，我们采取了分阶段从 DDR3 内存中调取数据的方案，确保了 PL 端能够更好执行卷积、激活和池化等操作。

PS 与 PL 的交互使用 AXI 接口协议使用 DDR3 进行数据交互，PL 通过 AXI_Lite 接口接收 PS 端的控制指令和各种参数，然后利用其硬件加速能力快速处理图像数据，完成后通过中断信号通知 PS 端。AXI_DMA 控制器在图像数据的传输中负责将图像数据从 DDR3 内存传输至 PL 端，以便进行卷积、权重等运算。为了确保卷积运算的顺利进行，所需的参数被预先保存在 SD 卡中，并在运算开始前加载到 PL

端的存储中，为高效的图像处理提供数据基础。

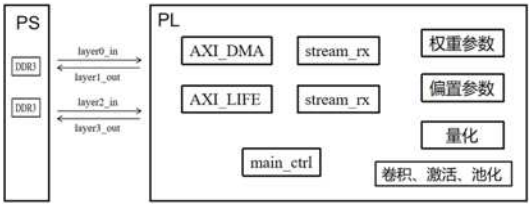

图 1 PS 与 PL 的交互

2.2 卷积运算设计

在深度学习模型中，卷积运算是核心且计算密集的步骤。由于 FPGA 的计算资源限制，无法同时计算大量数据，我们设计每次只计算 8 个通道的数据。简化了硬件资源的管理和分配。

以 YOLOv3 模型中的 layer2 为例，输入数据包含 16 个通道。首先调入前 8 个通道进行卷积运算，待 8 个通道的计算完成后，结果被暂存于 PL 端的 RAM 中。等剩下 8 个通道数据进行计算完后将其相加，获得完整的输出数据。我们通过分次处理解决输入数据超出 PL 端 RAM 存储能力的情况。先传输 10 行数据进行计算，并将这 10 行的计算结果存储于 PL 端 RAM 中。完成后再传输剩余的行数据，继续执行卷积运算。对于通道数不足 8 个的 layer，如 layer0，我们将前 3 个通道的数据将正常传输并参与运算，剩余的通道则通过补零的方式进行处理，保证整个卷积运算的连贯性和一致性。

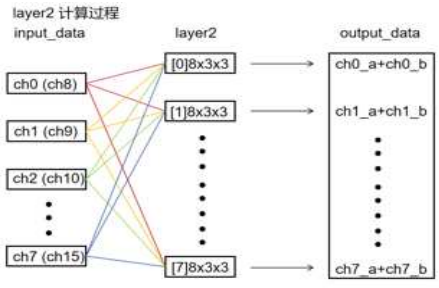

图 2 layer2 层计算过程

2.3 DSP48 完成卷积乘法

卷积操作是把输入的图像数据和卷积的权重参数先进行相乘，再把相应的结果再相加。在卷积中实现乘法操作输入的数据是 8bit × 8bit，而 DSP48 中的乘法操作为 25bit × 18bit，直接使用乘法器 IP：Multiplier 中的 DSP48 来说会对资源会造成浪费。为了提高 DSP48 的使用效率，对 DSP48 进行一个复用操作，实现两个 Int8 乘法，使用条件需要有两个 Int8 乘法具有相同因子。

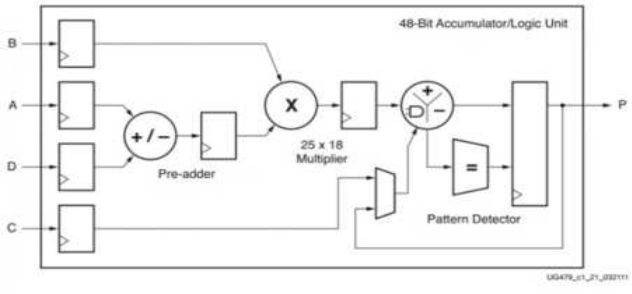

图 3 DSP48E1 结构图

使用原理：计算过程等效于（a+d）×b，将 A 端口 25bit 的高 9 位写入数据 a（a 的符号位有两位），低 16 位置零。将 D 端口 25bit 的低 9 位写入数据 d（和 a 一样对符号位进行扩展），高 16 位都置为 d 的符号位。B 端口 18bit 低 8 位写入数据 b，高 10 位都置为 b 的符号位，最后由 P 端口输出 ab（或 ab−1）和 db。

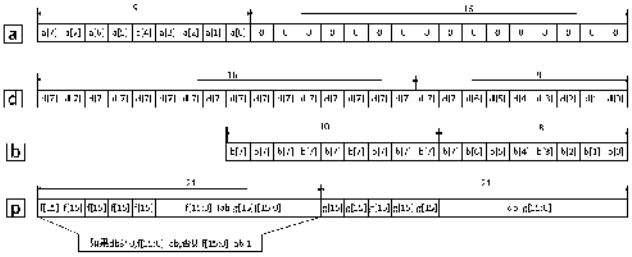

图 4 复用计算结构图

2.4 多通道卷积运算

运行卷积计算时，将 8 个通道的输入数据依次输入卷积通道里进行卷积运算，得出对应的输出数据。第一个输入通道中，与相应的 8 个权值进行相乘，用其中两个权值去实现 DSP48 的复用方案，则可以得出两个输出通道卷积相乘的运算结果，最后再进行相加得到第一个输出通道的结果。用 8 个 DSP48，输出 8 通道数据；2 个输出通道的权重，得到 2 个输出通道的卷积结果。将模块例化 4 次，就可以得到 8 个通道得输出结果。

2.5 显示方案介绍

在本系统中，我们使用了 OV5640 摄像头作为图像采集的硬件设备，该摄像头能够以 1280 × 720 的分辨率捕捉图像，并通过 HDMI 接口将图像清晰地展示在显示器上。采集到的图像数据将被分为两部分，有不同的处理流程和显示。

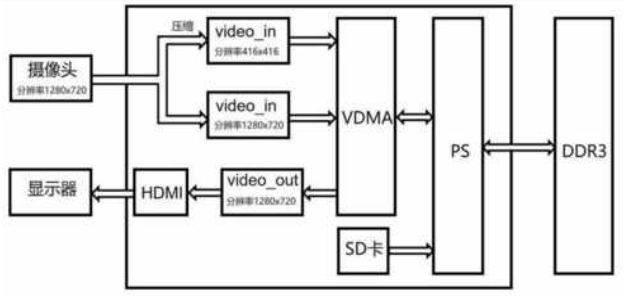

图 5 显示方案的架构

第一部分图像数据将原始图像尺寸调整为 416 × 416 像素通过 VDMA 通道传输至 PS 端，存储于 DDR3 内存中。PL 端从 DDR3 读取数据并执行卷积运算，识别出图像中的目标对象，将数据显示在显示器上。第二部分图像数据保持原始的 1280 × 720 分辨率作为直接显示的图像。

3 实验分析

3.1 显示效果

本次使用该系统运行后的结果如图 6 所示。

图 6 FPGA_YOLO 检测效果

图片 6 中展示出了本系统识别出来的各类物品并进行了标注。标注结果能较好得将物品识别出来。

3.2 性能分析

基于 ZYNQ-7030 实现得硬件加速 CNN 目标检测系统，逻辑资源消耗和功耗如图 7 所示。

Resource	Utilization	Available	Utilization...
LUT	13512	78600	17.19
LUTRAM	2119	26600	7.97
FF	16370	157200	10.41
BRAM	137	265	51.70
DSP	296	400	74.00
BUFG	1	32	3.13

图 7 逻辑资源消耗

在图 7 中可以看出系统在 BRAM 和 DSP 中消耗得资源较大，这一部分主要作用于图像得传输和卷积计算。在整个功能实现过程中，主要由 PS 端与 PL 端互通并与 RAM 进行数据交互，动态功耗较多。

4 总结

本文使用了基于 ZYNQ-7030 型号的 FPGA 开发板实现了硬件加速 CNN 目标识别，PS 端作为调度中心，负责数据预处理和结果解析；PL 端执行计算密集型操作，设计了基于 DSP48 的多通道卷积运算，展示了 FPGA 在深度学习目标识别任务中的潜力，实现了高效的数据处理和低功耗运行，为嵌入式系统提供了一定的技术支持。未来的工作可以进一步优化资源消耗和功耗，以及探索更复杂的模型和更高效的运算策略。

参考文献：

[1] QIUJ,SONG S,WANG Y,et al.Going deeper with embedded FPGA platform for convolutional neural network [C] //the 2016 ACM/SIGDA International Symposium,ACM,2016:26-35.

[2] 李 林，张盛兵，吴 娟 . 基于深度学习的实时图像图像目标检测系统设计 [J]. 计算计算机测量与控制，2019，27（7）:15-19.

[3] GAREA A S, HERAS D B, ARGUELLO F. Caffe CNN based classification of hyperspectral images on GPU[J].The Journal of Supercomputing, 2019,75(3):1065-1077

[4] LI Z G,WANG JT. An improved algorithm for deep learning YOLO network based on Xilinx ZYNQ FPGAC[C]//2020 International Conference on Culture-oriented Science & Technology(ICOST). Beijing: IEEE, 2020:447-451.