DOI:10.19651/j.cnki.emt.2210454

基于 FPGA 与退化 YOLO 的手机镜片缺陷检测系统*

王习东[1] 王国鹏[1] 王保昌[2] 张 浩[2] 冯文杰[1] 杨业泉[3]
(1.三峡大学计算机与信息学院 宜昌 443002; 2.三峡大学理学院 宜昌 443002; 3.三峡大学电气与新能源学院 宜昌 443002)

摘 要: 针对镜片缺陷检测采用图像处理法和神经网络法存在时延高、功耗高和检测缺陷类别较少等问题,设计了一种基于 FPGA 与退化 YOLO 的软硬协同检测系统。系统中使用卷积层代替 YOLO 网络的重排序层进行网络退化,并映射到 FPGA 上;采用动态量化、模块融合、双缓冲流水线、循环展开和分块等优化策略,设计可动态配置的加速 IP,其中的卷积计算模块分别实现了基于 Winograd 和 GEMM 的快速卷积算法。实验结果表明,本系统的加速 IP 在 PYNQ-Z2 上获得了 51.89 GOP/s 的计算性能,比基于典型滑动窗口卷积计算方法的性能提高了 0.76 倍,加速单张图像的时延为 433 ms,功耗为 1.07 W,与 Core i5-10500 CPU 相比,能效是其 365.27 倍,实现了小型设备对手机镜片低时延、低功耗的多缺陷检测。

关键词: FPGA;YOLOv2;手机镜片检测;软硬协同检测;快速卷积算法
中图分类号: TP391 **文献标识码:** A **国家标准学科分类代码:** 510.4030

Mobile phone lens defect detection system based on FPGA and degraded YOLO

Wang Xidong[1] Wang Guopeng[1] Wang Baochang[2] Zhang Hao[2] Feng Wenjie[1] Yang Yequan[3]
(1. College of Computer and Information Technology, China Three Gorges University, Yichang 443002, China;
2. College of Science, China Three Gorges University, Yichang 443002, China;
3. College of Electrical and New Energy, China Three Gorges University, Yichang 443002, China)

Abstract: Aiming at the problems of high delay, high power consumption and less defect categories in lens defect detection using image processing method and neural network method, a software and hardware collaborative detection system based on FPGA and degraded YOLO is designed. In the system, the convolution layer is used to replace the reordering layer of the YOLO network for network degradation and mapped to FPGA; dynamic quantization, module fusion, double buffer pipeline, loop expansion and block segmentation optimization strategies are adopted to design dynamically configurable acceleration IP. The convolution calculation module implements fast convolution algorithms based on Winograd and GEMM respectively. The experimental results show that the acceleration IP of this system obtains the calculation performance of 51.89 GOP/s on PYNQ-Z2, which is 0.76 times higher than that based on the convolution calculation method of typical sliding window. The time delay of the acceleration single image is 433 ms, and the power consumption is 1.07 W, compared with Core i5-10500 CPU, the energy efficiency is 365.27 times higher, which realizes the multi-defect detection of low delay and low power consumption of mobile phone lens by small equipment.

Keywords: FPGA; YOLOv2; mobile phone lens detection; software and hardware collaborative detection; fast convolution algorithm

0 引 言

随着卷积神经网络在目标检测[1]、智慧交通[2]、医学图像识别[3]和金融量化分析[4]等领域的蓬勃发展。当前数据中心、服务器端在处理卷积神经网络大批量数据时,考虑到能效等因素,会大规模部署可编程门阵列(field programmable gate array, FPGA)。FPGA 也可以很好地满足嵌入式端低功耗高性能的需求[5],并且其动态可重构

收稿日期:2022-06-24
* 基金项目:国家自然科学基金面上项目(52179136)资助

的特点能完美适应不断更新迭代的卷积神经网络模型[6]。

手机光学镜片由于制造工艺复杂,精度要求高,镜片生产过程中不可避免地出现各种各样的缺陷。因此,确保镜片质量的最有效措施是严格依照镜片质量要求进行镜片缺陷检测。目前国内企业主要采用人工目视法检测镜片缺陷。该方式效率低、误检率高、易受主观因素的影响,未适应当下的自动化趋势。常规的机器视觉法有两种方案,一种是传统的图像处理:孙力等[7]采用区域生长、形态学运算和椭圆拟合等算法,结合 Sobel 边缘检测、阈值分割方法实现对水印缺陷的识别与定位;曹宇等[8]通过改进(PSO)+(Otsu)权重因子更新策略,计算粒子的最优位置,最终实现光学镜片图像的阈值分割。另一种是基于神经网络:孟奇等[9]提出一种基于双通道生成对抗网络的数据增强方法,使用双通道鉴别器分别判断生成的整体缺陷图像和缺陷目标的真实性,对缺陷目标的类别进行鉴别;王国鹏等[10]利用 YOLOv2 网络模型,提出一种手机镜片缺陷实时检测方法。传统的图像处理算法,主要为滤波和提取边缘信息,难以实现多缺陷检测。基于神经网络的方法,大部分能实现多缺陷检测,但普遍用于服务器端和高性能嵌入式设备端,存在检测时延高、功耗高、成本高和部署条件苛刻等不足。因此亟需一种能在小型设备上满足低时延、低功耗、高准确率和强稳定性需求的镜片缺陷检测方法,助力企业降低人力成本和工人的劳动强度。

本文搭建了基于 XY 平台、图像采集装置、HDMI 显示器和 FPGA(PYNQ-Z2)等硬件为基础的手机镜片缺陷检测系统;在 PYNQ-Z2 平台上部署了一种退化 YOLO 网络模型,并通过高层次综合(high-level synthesis, HLS)设计了基于动态量化、模块融合、双缓冲流水线、循环展开和分块等优化机制的软硬协同检测方法。本文重点研究内容如下:1)针对 YOLOv2 中重排序层的切片(Slice)操作加重数据处理负担,使用卷积层代替重排序层进行网络退化,降低缓存的占用;2)采用动态定点 16 位量化对模型的权重和偏置参数、输入与输出特征图和中间结果进行量化,降低计算复杂度和硬件资源消耗;3)融合批归一化层的参数至卷积层,减少计算量和数据传输量;4)对数据的输入与输出设计双缓冲存储流水线机制,提高传输性能和数据吞吐量;5)针对网络中计算开销大且计算结构冗余的 3×3 和 1×1 卷积层分别实现了基于 Winograd 和通用矩阵乘(general matrix multiplication, GEMM)的快速卷积算法,提高卷积计算的性能。通过手机镜片缺陷检测实验,验证了本系统的有效性和实用性。

1　退化 YOLO 网络加速 IP

1.1　退化 YOLO 网络模型

尽管 YOLO 系列模型[11-13]在目标检测领域表现很好,但是并非一套网络框架能适用于所有,针对具体的手机镜片缺陷检测问题,需要设计相应的方法和结构。以目前在工业界广泛应用的 YOLOv2 为基础模型,来构造在 FPGA 平台上部署的退化 YOLO 网络模型,其必要性如下:1)YOLOv2 使用轻量级的 darknet-19 框架作为特征提取骨干网络,该框架是由 C 语言编写的,可通过 HLS 快速地描述并实现成对应的硬件架构,网络移植性好;2)手机镜片缺陷特征大多为线条、点状、絮状等简单特征,选用对小尺寸目标检测精度高、实时检测速度快的 YOLOv2 模型进行网络退化,即可达到较好的检测效果;3)PYNQ-Z2 的成本较低,计算资源相对有限,综合考虑硬件检测性能和模型部署难度等方面,应用在 YOLOv2 上较为合适。

重排序层的 Slice 操作如图 1 所示,Slice 操作类似于下采样取值,把大尺寸的特征图分成多张小尺寸的特征图,将空间信息绕到了特征通道中,保持了下采样取值信息没有丢失。作者设计的初衷是为了减少浮点数和加快运算,但是特征图数量的增加,对于芯片,特别是不含图形处理器(graphics processing unit, GPU)、神经网络处理器(neural network processing unit, NPU)加速的芯片,Slice 操作只会让缓存占用严重,加重计算处理的负担。同时,在芯片部署阶段,重排序层的转化将占用较多的逻辑资源,时序控制的要求也很高。经过试验比较,使用卷积层代替重排序层进行网络退化,能降低缓存的占用,提高检测性能,且更易部署。

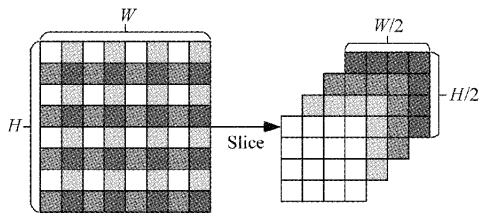


图 1　重排序层的 Slice 操作示意图

使用卷积层代替重排序层后的退化 YOLO 网络模型结构如图 2 所示,模型共包含两类组件:CBL 和 MCN。其中 CBL 为退化 YOLO 网络结构中的最小组件,由 Conv(卷积运算)、BN(批归一化)和激活函数 Leaky ReLU 组成。MCN 为降采样组件,由最大池化层 Max 和 N 个 CBL 构成。

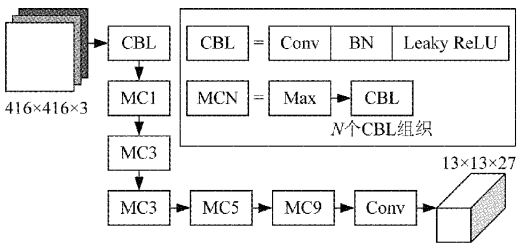


图 2　退化 YOLO 网络模型结构

退化 YOLO 模型另外两点变化如下:1)预测边界框由 5 个减少至 3 个。手机镜片缺陷检测中一般 2 个预测边界框就能完全框住缺陷,使用 3 个边界框,选择与真实标定框

交并比(intersection over union，IoU)更大的边界框，能更准确的框住缺陷。2)剔除非极大值抑制机制。手机镜片缺陷特征尺寸较小，基本上都是完整的落在一个网格内，不存在跨网格被多次检测的情况，无需清除被重复检测的结果。

1.2　加速 IP 框架

加速方案将退化 YOLO 网络模型中计算量大且计算结构冗余的部分植入 FPGA 中，在可编程逻辑(programmable logic，PL)端，设计可动态配置的加速 IP(intellectual property)，处理系统(processing system，PS)端存储退化 YOLO 网络的结构参数，通过调用加速 IP 的方式，得到计算结果。

加速 IP 的总体框架如图 3 所示，在 FPGA 的 PL 端，设计数据输入与输出模块，将卷积计算和最大池化计算进行 IP 模块封装，封装 AXI-lite 和 AXI-full 接口，PS 端通过 AXI-lite 接口对计算单元进行配置，通过 AXI-full 接口将权重参数和样本特征数据传入加速 IP 中，再由 AXI-full 接口接收计算的结果。

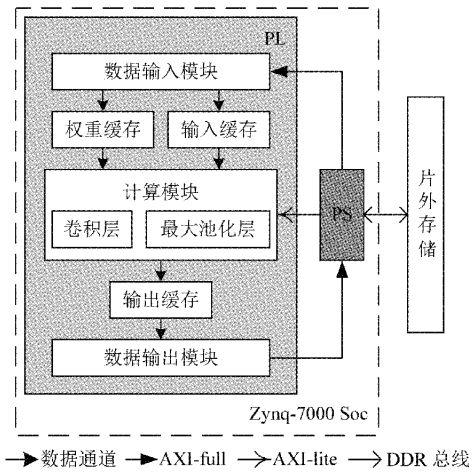


图 3　退化 YOLO 网络加速 IP 框架

1.3　动态定点 16 位量化

未经量化前，数据形式为浮点 32 位，对传输总线的带宽要求高。本文采用动态定点 16 位量化[14]对模型的权重和偏置参数、输入与输出特征图和中间结果进行量化，降低数据的位宽，减少计算量和数据传输量[15]。遍历寻找最优阶码，如式(1)所示。

$$exp_Q = arg\min \sum_{i=0}^{n} \left| Q_{float}^{i} - Q_{(bw,\exp_w)}^{i} \right| \tag{1}$$

式中：exp_Q 为量化后精度损失最小的阶码，n 为需量化的数据量，Q_{float}^{i} 为第 i 个数的原始浮点数值，$Q_{(bw,\exp_w)}^{i}$ 为第 i 个数在位宽 bw 和阶码 exp_Q 下，先量化成定点数，再转换回新的浮点数值。

1.4　批归一化层融合至卷积层

常规的前向推理过程，需依次计算卷积层和批归一化层。卷积层的计算如式(2)所示。

$$\boldsymbol{x}_{conv} = \boldsymbol{w}_i \otimes \boldsymbol{x}_i + \boldsymbol{b} \tag{2}$$

式中：$\boldsymbol{x}_{conv}$ 为卷积后结果，$\boldsymbol{w}_i$、$\boldsymbol{x}_i$ 和 $\boldsymbol{b}$ 分别为卷积核的权值、输入特征图和偏置。批归一化层的计算如式(3)所示。

$$\boldsymbol{x}_{bn} = \boldsymbol{\gamma} \frac{\boldsymbol{x}_{conv} - \boldsymbol{\mu}}{\sqrt{\sigma^2 + \epsilon}} + \boldsymbol{\beta} \tag{3}$$

式中：$\boldsymbol{x}_{bn}$ 为批归一化后结果，$\boldsymbol{\gamma}$ 和 $\boldsymbol{\beta}$ 分别为批归一化计算的缩放因子和偏置，$\boldsymbol{\mu}$ 和 σ^2 分别为特征图的均值和方差，ϵ 为正则化参数。

将式(2)代入式(3)中，可得批归一化参数融合至卷积的计算形式，如式(4)所示。

$$\boldsymbol{x}_{bn} = \boldsymbol{\gamma} \frac{(\boldsymbol{w}_i \otimes \boldsymbol{x}_i + \boldsymbol{b}) - \boldsymbol{\mu}}{\sqrt{\sigma^2 + \epsilon}} + \boldsymbol{\beta} = \frac{\boldsymbol{\gamma}\boldsymbol{w}_i}{\sqrt{\sigma^2 + \epsilon}} \otimes \boldsymbol{x}_i + \frac{\boldsymbol{\gamma}(\boldsymbol{b} - \boldsymbol{\mu})}{\sqrt{\sigma^2 + \epsilon}} + \boldsymbol{\beta} = \boldsymbol{w}_{new} \otimes \boldsymbol{x}_i + \boldsymbol{b}_{new} \tag{4}$$

式中：由原参数计算得出的新权值 $\boldsymbol{w}_{new}$ 和新偏置 $\boldsymbol{b}_{new}$ 作为新的卷积计算参数，可直接得出卷积后的批归一化结果。在 FPGA 加速卷积计算前，融合批归一化层的参数至卷积层，可减少模型计算量和数据传输量，加快模型推理速度。

1.5　数据输入与输出模块

数据输入与输出模块作为主器件，其接口设计采用 AXI-full 接口，PS 端的 DDR(double data rate)作为从器件，考虑到 FPGA 的 BRAM(block random access memory)资源有限，将采用循环分块的思路进行展开，如式(5)和(6)所示。

$$T_{ci} = (T_{co} - 1) \times S + K \tag{5}$$

$$T_{ri} = (T_{ro} - 1) \times S + K \tag{6}$$

式中：T_{ci} 和 T_{ri} 为输入特征图的大小，T_{co} 和 T_{ro} 为输出特征图的大小，S 为卷积核的步长，K 为卷积核的大小。

数据输入与输出模块在深度上进行展开的过程如图 4 所示，T_n 为输入深度，T_m 为得到的输出深度。采用双 Buffer 的缓冲设计，并进行流水线操作，加大数据传输的吞吐量，避免对后续的卷积计算加速模块造成性能瓶颈。

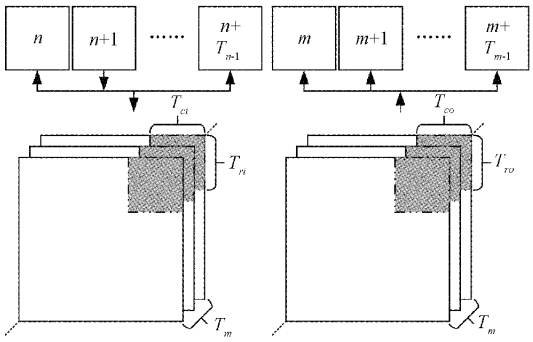


图 4　输入与输出分块示意图

双 Buffer 缓冲的流水线输入与输出结构如图 5 所示。图 5 若为流水线输入结构，则 A 为将数据从 DDR 传入缓存 Buffer1，a 为将数据从 DDR 传入缓存 Buffer2，B 为将数据从 Buffer1 传入片上输入缓存，b 为将数据从 Buffer2 传

入片上输入缓存;图 5 为流水线输出结构,则 A 为将数据从片上输出缓存传入缓存 Buffer1,a 为将数据从片上输出缓存传入缓存 Buffer2,B 为将数据从 Buffer1 传入 DDR,b 为将数据从 Buffer2 传入 DDR。

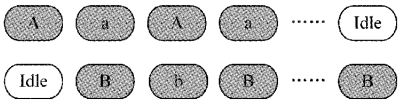


图 5　双 Buffer 缓冲的流水线输入与输出结构

1.6　卷积计算模块

为提高卷积计算并行度,采取展开和分块这两种优化策略。对输入和输出特征图进行二维展开,每次并行计算 T_m 个输出特征图和 T_n 个输入特征图,并将输入的特征图进行分块,每次仅对一块进行计算。降低 FPGA 对 BRAM 的需求,复用计算数据,减少从片外存储读取和写入数据的次数。

1)Winograd 快速卷积计算模块

针对退化 YOLO 网络中卷积核大小为 3×3 的卷积计算,采用基于 Winograd 的快速矩阵乘法计算引擎。Winograd 算法分别对输入特征图和卷积核变换成相同维度的矩阵,然后变换矩阵通过点乘来完成卷积计算,如式(7)所示。

$$\boldsymbol{O}=\boldsymbol{X}^{\mathrm{T}}[[\boldsymbol{ZKZ}^{\mathrm{T}}]\odot[\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{IY}]]\boldsymbol{X}=\boldsymbol{X}^{\mathrm{T}}[\boldsymbol{K}'\odot\boldsymbol{I}']\boldsymbol{X} \tag{7}$$

式中:$\boldsymbol{O}$ 为卷积后的输出特征图,$\boldsymbol{I}$ 为输入特征图中单通道的子特征矩阵,$\boldsymbol{K}$ 为卷积核单个通道上的元素,$\boldsymbol{X}$、$\boldsymbol{Y}$、$\boldsymbol{Z}$ 及其对应的转置为 Winograd 算法中的转换系数矩阵。

对于算法规模为 $F(2,3)$的典型滑动窗口卷积计算,需经过 6 次乘法和 4 次加法,而 Winograd 快速卷积计算能减少 2 次乘法,但增加了 4 次加法。Winograd 算法减少乘法次数带来的性能提升远高于增加加法次数所损失的性能。

本文通过 HLS 对 Winograd 算法进行硬件架构设计,首先使用数组分割指令将不同通道上的元素分割,实现不同行上的缓存数据并行读取;然后指定寄存器类型,避免矩阵变换时造成的访问冲突,同时减少矩阵点乘所需的时钟周期数;最后使用控制指令对输出通道上的缓存数据进行循环展开和流水线工作。

基于 Winograd 快速卷积计算模块设计如图 6 所示,输入特征图各通道上的并行卷积过程为,首先从输入缓存和卷积核缓存中并发读取特征图和卷积权值,在各通道上将输入矩阵和权值矩阵分别变换成相同的矩阵。然后对变换结果并行计算矩阵点乘以及变换输出矩阵。最后经过并行加法树结构计算输出特征图在各通道上的卷积结果。

2)GEMM 快速卷积计算模块

针对退化 YOLO 网络中卷积核大小为 1×1 的卷积计算,采用基于 GEMM 的快速矩阵乘法计算引擎。GEMM 算法配合 Im2col 操作对输入特征图的通道数据和卷积核的窗口数据扩展成列,通过多级存储结构和程序执行的局部性原理,提高卷积的计算效率。

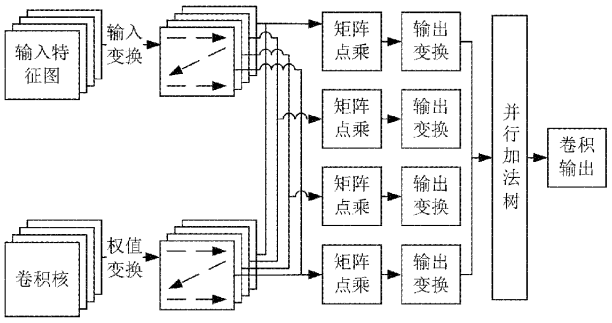


图 6　Winograd 快速卷积计算模块

本文通过 HLS 对 GEMM 算法进行硬件架构设计,采用循环分块策略,对卷积内循环实现基于 GEMM 快速卷积计算,该模块设计如图 7 所示。卷积内循环中输入特征图各通道上的并行卷积过程为,首先对输入特征图缓存中的特征值与卷积核缓存中的卷积权值,并行计算矩阵点乘。然后经过并行加法树结构计算输出卷积结果。

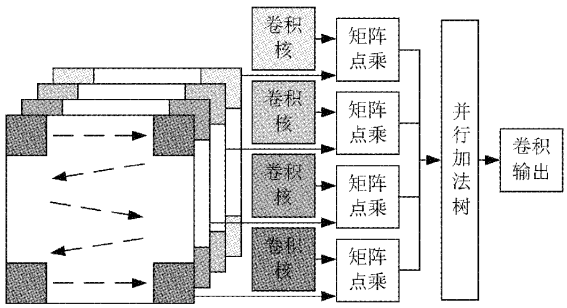


图 7　GEMM 快速卷积计算模块

1.7　最大池化计算模块

由于 Winograd 的卷积输出窗口与池化窗口的滑动规律完全相同,为避免单独设计最大池化计算模块引入的传输延时,将最大值池化计算与卷积计算融合,把最大池化看成是一种特殊的卷积。它不需要卷积权值,只对输入特征图在单通道上进行最大元素输出,并且它所对应的运算单元是比较器。

最大池化计算模块设计如图 8 所示,融合后的计算过程为,寄存器初始置零,当卷积激活后的值依次输入时,比较器将大的数输入到寄存器中完成比较,将最大的数输出,同时将寄存器清零。由于比较器消耗的资源多为查找表(look-up-table, LUT),最大池化设计采用分块的优化策略。

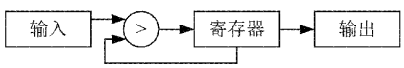


图 8　最大池化计算模块

2　手机镜片缺陷检测系统

2.1　系统整体组成

本文实现的手机镜片缺陷检测系统主要由 FPGA、XY

平台模块、图像采集模块、HDMI 显示模块组成，整体结构如图 9 所示，系统主控器为 FPGA 的 PS 端，其主要功能包括存储退化 YOLO 网络结构、驱动 XY 平台、调用工业相机、配置加速 IP、预处理采集的图像和对加速 IP 最后的输出进行后处理。实际搭建的系统如图 10 所示。

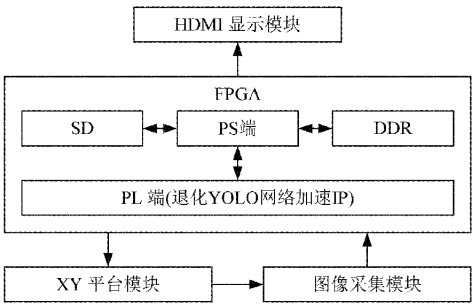


图 9　系统整体结构

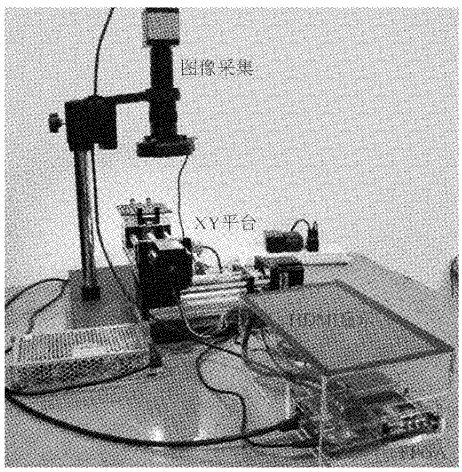


图 10　实际搭建的系统

2.2　系统模块功能

1)XY 平台模块

XY 平台传动装置的主体为两个小型步进电机十字形组合而成，实现手机镜片在二维水平面上的移动与定位。XY 平台模块由 FPGA 的 PL 端，通过通用输入输出端口 GPIO(general-purpose input/output)向外部引脚发送特定的脉冲波形控制电机步进移动，电机侧面安装的限位开关通过向 PL 端发送归零电平信号完成镜片相对位置的初始化。

2)图像采集模块

图像采集模块由工业相机、显微镜头和升降支架台组装而成，PS 端通过调用 OpenCV 驱动工业相机，其内置函数会将视频多缓存 4 帧，为解决连续采集造成视频非实时帧的问题，编写跳帧和释放多余缓存帧的程序段。

3)HDMI 显示模块

镜片缺陷检测结果采用 HDMI 进行显示，使用 DIGILENT 公司的 RGB2DVI 开源 IP 核完成数据位转换，使用 XILINX 公司的 VTC IP 核控制视频时序。为了提高视频传输速率和简化 PS 端的工作量，在 PL 端调用 VDMA IP 核，通过 AXI-full 读取 DDR 的数据并将数据转换为 AXI-stream 格式，调用 AXIS2VIDEO IP 核，将 AXI-stream 总线传进来的数据转换成 24 位的 RGB 数据并同时将从 VTC IP 核接收的视频时序，输出给 RGB2DVI IP 核。AXIS2VIDEO IP 核与 VDMA IP 核的数据交互通过 AXI-stream 总线进行。

为减少显示实时检测结果的耗时，采用 ipywidgets 库实现，将处理后的图像直接显示，若图像上存在缺陷则保持存储，若无任何缺陷，则直接覆盖为下一帧。

2.3　软硬协同检测

1)PYNQ 框架

开源 PYNQ(python productivity for Zynq)框架，即在原有的 Zynq 架构上，添加了对 python 的支持，将 ARM 与 FPGA 的底层交互逻辑进行封装，借助 python 语言丰富的第三方扩展库，FPGA 被抽象成若干个加速 IP，PS 端通过执行 python 脚本动态加载 bitstream，DMA(direct memory access)将数据流传输到加速 IP 融合输出，高效开发 PL 端的存储和计算资源，充分发挥 FPGA 软硬协同处理的优势[16]。

2)PS 与 PL 端协同检测

FPGA 的 PS 与 PL 端软硬协同实时检测手机镜片缺陷的流程如图 11 所示。首先 FPGA 的 PS 端将存储在 SD 卡中经动态定点 16 位量化后的权重和偏移参数文件读入 DDR 片外存储器；接着驱动 XY 平台步进移动，并调用工业相机采集手机镜片实时图像并进行预处理，主要包括采集图像的整形和像素归一化；然后读取退化 YOLO 模型的网络结构，根据输入与输出的层数和阶码的值，配置 PL 端的加速 IP，根据分工对卷积层、最大池化层进行加速；最后

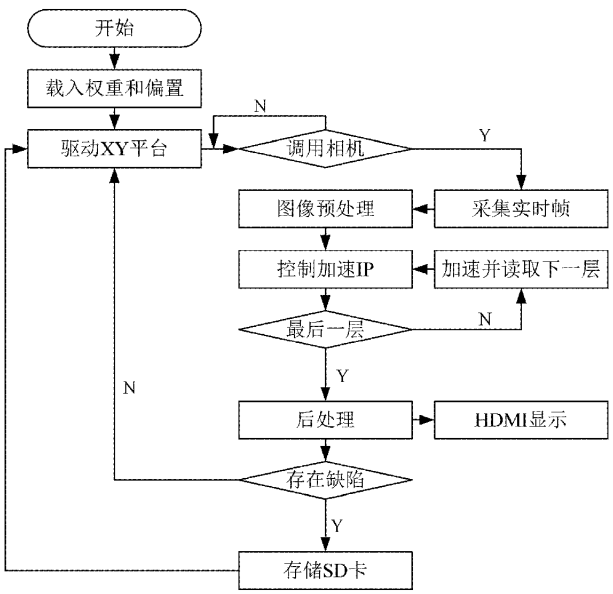


图 11　软硬协同实时检测流程

对加速 IP 的输出进行后处理,主要包括前向推理过程最后的检测层和镜片缺陷预测框的定位及类别的标定,将检测结果通过 HDMI 输出显示,若镜片存在缺陷,则将缺陷检测结果存储至 SD 卡中。

3　实验及评估

3.1　手机镜片缺陷检测数据集

选取手机镜片常见的 4 类缺陷(花瓣伤、胶洞、裂痕、膜裂)为研究对象制作数据集。训练与验证需要大量样本,本文通过数据增强(旋转变换、镜像变换、图像增强和添加噪声等)的方式,扩充样本数量。这样既增加了样本数量,又变换了手机镜片缺陷的方向和大小,还增强了样本图像噪声的鲁棒性[17]。手机镜片缺陷检测数据集划分如表 1 所示。

表 1　手机镜片缺陷检测数据集划分

参数	数量/张
总数据集	9 600
训练集	7 200
验证集	2 400

3.2　软硬件平台环境

检测系统使用的软硬件平台环境如表 2 所示。

表 2　软硬件平台配置

平台	配置
FPGA	PYNQ-Z2(Zynq7020)
框架	PYNQ
工具	Vivado 2020.2
语言	Python 3.7.4

3.3　训练过程曲线

训练过程中,退化 YOLO 网络模型的 Loss 变化曲线如图 12 所示。当训练迭代至 4 800 次时,损失函数收敛在 0.01 附近,表示模型收敛。

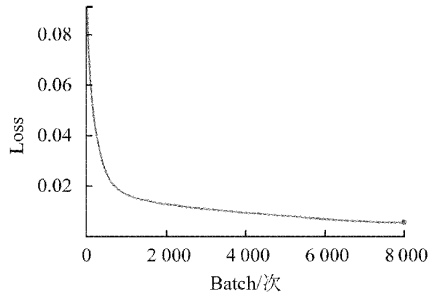


图 12　Loss 变化曲线

3.4　模型权重分离

改写特征提取网络框架源码,将模型权重分离为权重和偏置参数两个文件。映射如式(8)所示。

$$y = A \times X + B \tag{8}$$

式中:A 和 B 为映射系数。卷积核的权重和偏置参数表达式如式(9)和(10)所示。

$$weight_i = weight_i \times A_i \tag{9}$$

$$bias_i = bias_i \times A_i + B_i \tag{10}$$

式中:A_i 和 B_i 为特征图 i 的映射系数。将分离后的文件以二进制的形式存储映射后的值,减少计算量以提高运算速度[18]。

3.5　权重和偏置参数量化

首先将退化 YOLO 模型中批归一化层的参数融合至卷积层,然后采用动态定点 16 位量化融合后的权重和偏置参数文件。最终两个参数文件由浮点 32 位形式的 115.85 MB 量化为定点 16 位形式的 57.87 MB。并通过遍历寻找到了各卷积层权重和偏置参数的最优阶码,如图 13 所示。图 13 中,还有一条点折线表示的是各卷积层输入特征图的阶码,将该折线向右平移一个单位后即为该层输出特征图的量化阶码。

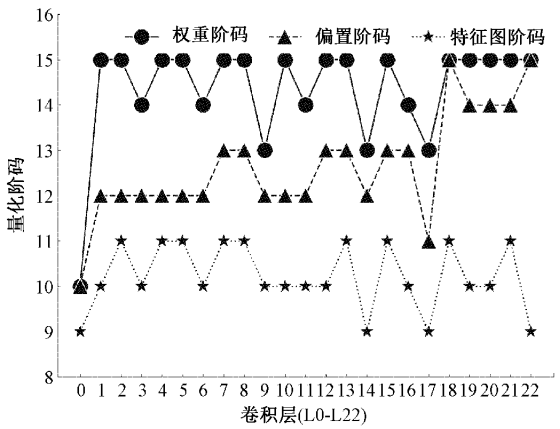
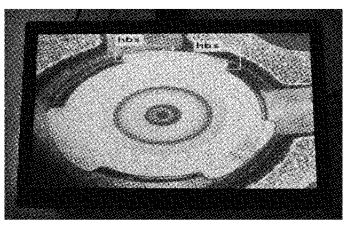


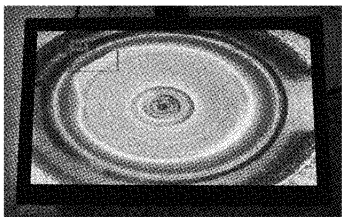
图 13　卷积层量化的最优阶码

3.6　实时检测

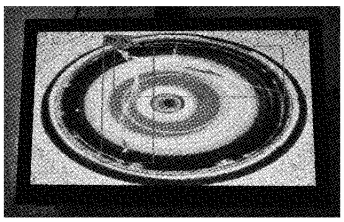
手机镜片缺陷验证集样本在 PYNQ-Z2 上实时采集图像进行检测,HDMI 显示检测结果如图 14 所示。图 14(a)模型正确框中 2 个花瓣伤缺陷,图 14(b)模型正确框中 1 个胶洞缺陷,图 14(c)模型正确框中 2 个裂痕缺陷,图 14(d)模型正确框中 1 个膜裂缺陷。目标框上方的数据为预测类别的标签。从检测结果来看,该量化模型能出色的预测并区分 4 类缺陷,同时在多缺陷共存的情况下,没有出现缺陷漏检情况。



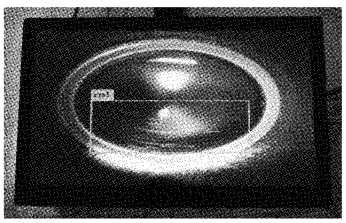
(a) 花瓣伤样本



(b) 胶洞样本



(c) 裂痕样本



(d) 膜裂样本

图 14　HDMI 显示实时检测结果

3.7　性能评估

1)检测准确率

手机镜片缺陷检测系统对验证集样本实时检测，统计验证结果如表 3 所示。检测准确率最低的手机镜片缺陷为花瓣伤；由于花瓣伤特征形态的尺寸较小，且与背景的差异也较小，导致其提取特征相对简单，检测准确率相对较低。

表 3　验证集的手机镜片缺陷检测结果

类别	个数/枚	准确数/枚	准确率/%
花瓣伤	1 152	1 004	87.15
胶洞	960	906	94.38
裂痕	192	190	98.96
膜裂	96	92	95.83

2)硬件资源消耗

将退化 YOLO 网络移植到 PYNQ-Z2 中，加速 IP 进行检测的硬件资源消耗表 4 所示。硬件资源消耗最多的为 LUT、数字信号处理器(digital signal processor，DSP)和 BRAM。为追求退化 YOLO 网络中大量乘加运算的加速效果最大化，将乘加运算都进行了并行展开，消耗了大量 LUT、DSP 资源。每次都要将权重参数和输入与输出特征图进行存储，消耗了大量 BRAM 资源。

3)性能与能效

本文设计的加速 IP，与未经量化的 CPU 平台和基于滑动窗口卷积实现的 YOLOv2 加速模型[19]，在性能与能效等方面的比较如表 5 所示。文献[19]的验证平台为 Zedboard，与本系统硬件部署的 PYNQ-Z2 片上总资源基本相同。分析表 5 可知，虽然采用动态定点 16 位量化相比于原浮点 32 位，在数据的值上会有一定的精度损失，但从模型的检测准确率均值看，本系统的加速 IP 和文献[19]都与原浮点 32 位模型的检测精度相比损失较小。从 FPGA 加速模型推理过程的功耗看，本系统的加速 IP 略低于文献[19]，主要是由于使用卷积层代替重排序层，降低了总的计算量。从单张图像的加速时延看，由于文献[19]对重排序层的 Slice 操作设计了单独的计算模块，因而会占用更多的缓存，增加数据读取和写入的时延；同时本系统的加速 IP 设计了多通道并行的 Winograd 快速卷积计算模块，在加速单张图像的时延上减少了 56.61%。从验证平台获得的计算性能和能效看，本系统的加速 IP 比基于典型滑动窗口卷积计算方法的性能提高了 0.76 倍，能效提高了 0.98 倍；与未经量化的 CPU 平台相比，性能是其 4.83 倍，能效是其 365.27 倍。

表 4　硬件资源消耗

资源类型	总量	消耗量	消耗率/%
LUT	53 200	28 000	52.63
LUTRAM	17 400	4 737	27.22
FF	106 400	33 646	31.62
BRAM	140	67	47.86
DSP	220	115	52.27

表 5　与 CPU 平台和其他相关研究的比较

参数	CPU	文献[19]	本文
验证平台	Core i5-10500	Zedboard	PYNQ-Z2
量化	Float-32	Fixed-16	Fixed-16
时钟频率/Hz	3.1 G	150 M	150 M
计算量/BFLOPs	29.35	29.35	22.47
准确率均值/%	96.13	94.70	94.08
功耗/W	81	1.20	1.07
加速时延/ms	2 729	998	433
性能/(GOP/s)	10.75	29.41	51.89
能效/(GOP/(s•W))	0.13	24.51	48.50

4　结　　论

为实现小型设备对手机镜片低时延、低功耗、高准确率和强稳定性的缺陷检测，本文构造了一种退化 YOLO 网络模型，并设计了基于动态量化、模块融合、双缓冲流水线、循环展开和分块等优化机制的软硬协同检测系统，其中的加速 IP 在片上时钟频率为 150 MHz 的 PYNQ-Z2 上获得了 51.89 GOP/s 的计算性能，比基于典型滑动窗口卷积计算方法的性能提高了 0.76 倍，加速单张图像的时延为

433 ms,功耗为 1.07 W,与 Core i5-10500 CPU 相比,能效是其 365.27 倍,验证了本系统在性能方面具有一定的优势,具有一定的应用价值。

后续考虑将检测系统部署在更高性能的 Zynq-Z 系列开发板,采用 32 位量化方法,控制检测精度与原浮点 32 位模型相一致,进一步提高检测性能。

参考文献

[1] 李明阳,胡显,雷宏. 基于可变形卷积神经网络的遥感图像飞机目标检测[J]. 国外电子测量技术,2020,39(7):121-126.

[2] 殷礼胜,孙双晨,魏帅康,等. 基于自适应 VMD-Attention-BiLSTM 的交通流组合预测模型[J]. 电子测量与仪器学报,2021,35(7):130-139.

[3] 王桂棠,林桢哲,符秦沈,等. 联合生成对抗网络的肺结节良恶性分类模型[J]. 仪器仪表学报,2020,41(11):188-197.

[4] PEI S, SHEN T, GU C, et al. 3DACN: 3D augmented convolutional network for time series data[J]. Information Sciences, 2020, 513:17-29.

[5] YU J C, GUO K Y, HU Y M, et al. Real-time object detection towards high power efficiency[C]. 2018 Design, Automation & Test in Europe Conference & Exhibition, 2018: 704-708.

[6] 陈桂林,马胜,郭阳. 硬件加速神经网络综述[J]. 计算机研究与发展,2019,56(2):240-253.

[7] 孙力,刘晨,姚红兵. 基于机器视觉的树脂镜片水印疵病检测[J]. 江苏大学学报(自然科学版),2018,39(4):425-430.

[8] 曹宇,徐传鹏. 一种改进阈值分割算法在镜片缺陷检测中的应用[J]. 激光与光电子学进展,2021,58(16):219-224.

[9] 孟奇,苗华,李琳,等. 基于双通道生成对抗网络的镜片缺陷数据增强[J]. 激光与光电子学进展,2021,58(20):356-364.

[10] 王国鹏,王习东,王保昌,等. 基于 YOLOv2 网络模型的手机镜片缺陷实时检测方法[J]. 自动化与信息工程,2021,42(5):28-32.

[11] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2016:779-788.

[12] REDMON J, FARHADI A. YOLO9000: Better, faster, stronger[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2017:6517-6525.

[13] REDMON J, FARHADI A. YOLOv3: An incremental improvement[C]. IEEE Conference on Computer Vision and Pat-tern Recognition, 2018:1-6.

[14] QIU J, WANG J, YAO S, et al. Going deeper with embedded FPGA platform for convolutional neural network[C]. The 2016 ACM/SIGDA International Symposium, 2016:26-35.

[15] ZHANG C, LI P, SUN G, et al. Optimizing FPGA-based accelerator design for deep convolutional neural networks[C]. The 2015 ACM/SIGDA International Symposium, 2015:161-170.

[16] 崔毅,殳国华,李丹. 基于 PYNQ 开发板的二值神经网络分类模型研究[J]. 电气自动化,2019,41(5):53-56.

[17] 贾振卿,刘雪峰. 基于 YOLO 和图像增强的海洋动物目标检测[J]. 电子测量技术,2020,346(14):84-88.

[18] 刘海军,岳英杰. 基于 PYNQ 平台的双目视觉测距系统设计[J]. 电子世界,2020(16):184-186.

[19] 陈辰,柴志雷,夏珺. 基于 Zynq7000 FPGA 异构平台的 YOLOv2 加速器设计与实现[J]. 计算机科学与探索,2019,13(10):1677-1693.

作者简介

王习东(通信作者),博士,讲师,主要研究方向为光电检测技术及系统、弱磁检测、FPGA 应用开发等。

E-mail:xdwang@ctgu.edu.cn

王国鹏,硕士研究生,主要研究方向为机器视觉、FPGA 应用开发。

E-mail: Minalinsky2333@qq.com