文章编号:1001-5078(2024)03-0466-07　　·图像与信号处理·

轻量化卷积神经网络红外目标识别性能分析与 FPGA 实现

王　戈,李江勇,杨德振,张子林,柴　欣
(中国电子科技集团公司第十一研究所,北京 100015)

摘　要:随着深度学习应用于计算机视觉,其数据量大、网络层结构复杂,在硬件部署中存在资源不足、延时高等成为关键问题,本文通过分析五种较有代表性轻量化网络的优缺点,提出一种将轻量化网络应用到红外目标检测领域的基于 MobileNet 的轻量化网络改进,并以 FPGA 为硬件载体实现。该网络使用 Tanh 激活函数替代原有激活函数并简化网络层数,以适应红外目标的特征提取,针对深度学习目标检测算法在硬件实现方面存在的数据量大,资源占用大,运算延时高等问题,采用 FPGA 进行硬件实现。实验表明,在 Xilinx Zynq-7020 XA 开发板上,设定时钟频率 100 MHz,输入图像大小为 640×512,改进后的 MobileNet 在保证原相同精度情况下实现 5.1 ms 每张图像。

关键词:轻量化网络;MobileNet;FPGA 实现;模型优化
中图分类号:TN491;TP391.41　　**文献标识码:**A　　**DOI:**10.3969/j.issn.1001-5078.2024.03.019

Infrared target recognition analysis and FPGA implementation based on lightweight convolutional networks

WANG Ge, LI Jiang-yong, YANG De-zhen, ZHANG Zi-ling, CHAI Xin
(The 11th Research Institute of CETC, Beijing 100015, China)

Abstract: With the application of deep learning in computer vision, its large amount of data, complex network layer structure, insufficient resources in hardware deployment and high delay have become key problems. This paper, by analyzing the advantages and disadvantages of five representative lightweight networks, proposes a lightweight network improvement based on MobileNet, which applies lightweight networks to infrared target detection field. FPGA is used as the hardware carrier. In this network, Tanh activation function is used to replace the original activation function and the number of network layers is simplified to adapt to the feature extraction of infrared targets. In view of the problems existing in the hardware implementation of deep learning target detection algorithm, such as large amount of data, large resource occupation and high calculation delay, FPGA is adopted for hardware implementation. The experiment shows that on Xilinx Zynq－7020 XA development board, the clock frequency is set to 100 MHz and the input image size is 640×512. The improved MobileNet can achieve each image of 5.1 ms with the same accuracy as the original one.

Keywords: Lightweight network; MobileNet; FPGA implementation; model optimization

收稿日期:2023-04-25

1 引 言

随着机器视觉的发展,当前需要处理的图像都存在背景复杂、图像尺寸不一致、目标大小分布不一致等问题,在红外场景下往往会出现漏记、错记的问题,导致精确度不高[1]。当前随着深度学习在计算机视觉兴起,研究者发现深度学习模型在计算资源、存储空间等方面存在巨大局限。针对深度研究所需的计算资源的限制,通常常规硬件(如 CPU、GPU)存在算力不足、能耗过大、延时高等问题。为了克服这些局限性,当前研究者普遍采用 FPGA(Field-Programmable Gate Array 现场可编程门阵列)来实现图像处理算法。FPGA 的并行性和流水化设计具有其他硬件所不具备的优势,这使得 FPGA 成为一种更具有效性和可扩展性的解决方案[2]。但是算法结构复杂、网络参数多任然是算法部署需要解决的问题,对此由此研究者们开始设计更加轻量级模型。轻量化网络就在于为这些边缘设备提供高效、快速、低功耗的深度学习模型。轻量化网络通过在模型设计、训练和推理等方面引入一些有效的轻量化技术,如深度可分离卷积、通道剪枝、低比特量化等,来降低模型的计算和存储需求[3]。这不仅可以让模型更好地适应边缘设备的资源限制,也可以提高模型的准确性和效率,促进人工智能技术在各行各业的应用和发展。因此,轻量化网络的研究意义是非常重要的,它为我们提供了一种新的深度学习技术解决方案,使得深度学习模型能够更加普及和广泛地应用于各种场景[4]。

本文将介绍五种具有代表性的轻量化网络,并通过实验比对选出网络模型更小,识别率最高的 MobileNet 网络,由于深度可分离卷积由于不考虑输入深度,在 MobileNet 网络基础改进并实验在红外场景下效果,最后以 FPGA 硬件实现。在移动端、嵌入式设备、自动驾驶等领域,深度学习模型需要满足计算量小、模型体积小、运行速度快等要求[5]。因此,研究轻量化网络可以让深度学习在这些设备上得到更广泛的应用[6]。此外,轻量化网络也能够在计算资源有限的情况下提高深度学习的效率,加速深度学习算法的研究和应用,相比于传统的 CPU 和 GPU 实现方式,FPGA 具有更高的并行度、更低的功耗和更高的灵活性,能够根据具体需求定制化设计,满足不同场景下的应用需求。

2 轻量化网络

2.1 SqueezeNet

SqueezeNet 是一种轻量级卷积神经网络,在保持相同精度的情况下,具有更小的模型大小和更少的计算资源消耗[7]。SqueezeNet 的核心思想是采用“Fire Module”来代替传统卷积神经网络中的“Convolutional Layer”。Fire Module 由一个称为“squeeze layer”的 1×1 卷积层和一个称为“expand layer”的具有不同卷积核大小的卷积层组成如图 1 所示。1×1 卷积层在不改变特征图大小的情况下,通过将通道数从输入通道数(即输入特征图的深度)压缩到较小的通道数,从而实现了参数和计算量的压缩。

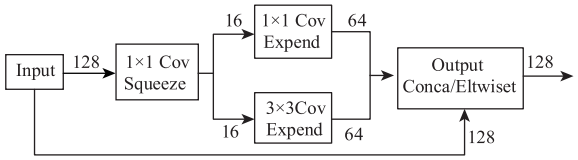


图 1 SqueezeNet 中的 fire 层结构

Fig. 1 Fire layer structure in SqueezeNet

在此之后,具有不同卷积核大小的 expand 层分别对压缩后的特征图进行卷积操作,从而增加了模型的非线性拟合能力。同时,为了减少模型的参数量,中采用了“Bypass connections”(即跳跃连接)将一些输入特征直接传递到输出特征中,这使得 SqueezeNet 成为一个轻量级卷积神经网络的代表。它在保持一定准确率的情况下,可以大幅减少模型的参数量和计算量,从而可以在保证性能的情况下,显著提高模型的速度和响应时间。

2.2 MobileNet

MobileNet 最早的研究是为了深度学习能够搭载在移动端,该网络在保持一定准确率的前提下,减少模型的参数量和计算量[8]。其中起到最关键的是 MobileNet 中的深度可分离卷积。一般的卷积需要采用对应输入通道的卷积进行,输出为每个通道对应卷积后的值,操作如图 2 所示。

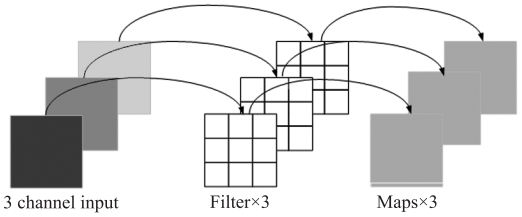


图 2 普通卷积

Fig. 2 General convolution

而 MobileNet 模型由若干个卷积层和深度可分离卷积层组成。深度可分离卷积是一种分解卷积运算的方法,将标准卷积分解为深度卷积和逐点卷积两个操作如图 3 与图 4 所示。相比于普通卷积,深度可分离卷积先用深度为 1 的 3 × 3 的卷积核(depthwise 分层卷积),再用 1 × 1 的卷积核(point-wise 卷积)调整通道数,将特征提取与特征组合分开进行深度可分离卷积,虽然每一层卷积变得更为复杂,但是通过该方法进行卷积在保证网络模型准确率的同时极大的减少了网络参数。

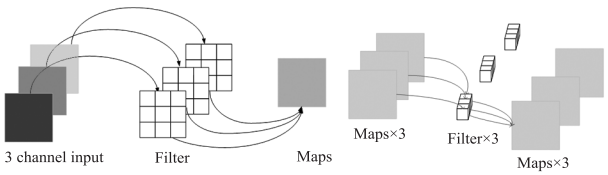


图 3 深度可分离卷积

Fig. 3 Depth-separable convolution

2.3 ShuffleNet

ShuffleNet 基本模块包括两个阶段,即分组卷积(group convolution)和通道重排(channel shuffle)。其中分组卷积是将输入特征图分为若干个组,并对每个组进行独立的卷积操作,该过程与深度可分离卷积方式相似,都是通过降低卷积的通道数来减少计算[9]。通道重排则将分组卷积后的输出特征图中的通道重新组合,从而增强不同组之间的信息流动。

ShuffleNet 整体结构由多个基本模块堆叠而成。其中,初始阶段:采用标准卷积对输入特征图进行初步的特征提取和下采样,减小特征图的尺寸,堆叠多个 ShuffleNet 基本模块,进一步提取特征并缩小特征图的尺寸,以解决信息流通不畅问题。采用 shuffle 替换掉 1 × 1 卷积,这样可以减少权值参数,而且是减少大量权值参数。整个网络的最后一层通常采用全局平均池化或全局最大池化操作,将特征图转换为向量。

分组卷积和通道重排有效地缓解了过拟合和梯度消失等问题,从而提高了网络的鲁棒性和泛化能力,可以方便地扩展到更深更复杂的网络结构中,从而适用于各种不同的任务和场景。当然该网络还有一些缺点,分组卷积和通道重排导致了网络的空间复杂度相对较高,可能会增加模型部署和存储的难度和成本,对于一些更深更复杂的网络,其精度仍有所限制,可能不适用于一些对精度要求较高的任务。

2.4 Xception 网络

Xception 是卷积神经网络的一种,该网络基于深度可分离卷积(Depthwise Separable Convolution)和 Inception 网络进行了改进。深度可分离卷积可以分离卷积操作的空间和通道维度,分别进行卷积操作[10]。这样做的好处是,可以大大减少计算量和模型参数数量,提高模型的计算效率。与 Inception 模块相似,通过并行的多个卷积核提取不同层次的特征,从而丰富特征信息,有利于提高模型的准确率。

Xception 网络相比于传统的卷积神经网络,Xception 中的深度可分离卷积可以大幅度减少计算量和模型参数数量,提高模型的计算效率。其次 Inception 模块的多个并行卷积核有利于提高模型的准确率。最后 Xception 网络具有较强的泛化性能,可以在各种任务中取得较好的性能表现。然而,Xception 网络也存在一些缺点,例如模型复杂度较高,训练时间和计算资源消耗较大。

2.5 EfficientNet

EfficientNet 是一种高效的卷积神经网络,由谷歌 AI 团队在 2019 年提出。它的设计思路是在保持准确性的同时,尽可能地减少计算复杂度和参数数量。网络采用一个复合的缩放方法,将深度、宽度和分辨率三个维度同时缩放,来构建更高效的神经网络[11]。具体地,首先从一个基本的网络结构开始,利用复合缩放系数对网络的深度、宽度和分辨率进行缩放,其中,深度通过增加网络层数来缩放,宽度通过增加通道数来缩放,分辨率则通过改变输入图像的大小来缩放。这样,通过复合缩放系数,EfficientNet 能够快速地找到在计算资源限制下,具有最高准确率的网络架构。

EfficientNet 在 ImageNet 数据集上取得了当时最好的结果,同时在计算复杂度和参数数量上都具有很高的效率。在相同准确率的情况下,与其他常用的卷积神经网络相比,EfficientNet 的计算复杂度和参数数量都要低得多。虽然 EfficientNet 已经算是比较轻量级,但是仍然可能会对计算资源造成一定的压力。在计算资源受限的情况下,推理的速度较慢可能会成为一个问题。

3　算法改进

3.1　实验分析

首先通过实验对比了五种种网络在 ImageNet 数据集上的表现，经 10 个 epoch 后分别对分类的 top1 和 top5 模型大小、网络深度进行统计，如表 1 中。

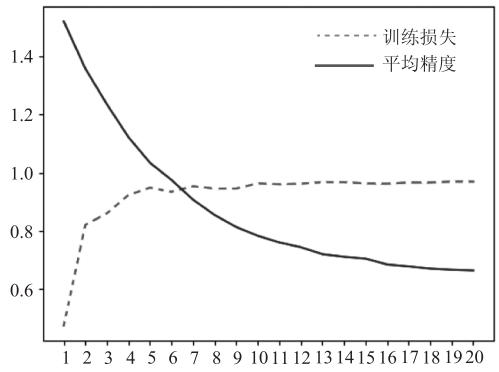
表 1　轻量化网络在 ImageNet 数据集上的表现

Tab. 1 Representation of lightweight networks on the ImageNet dataset

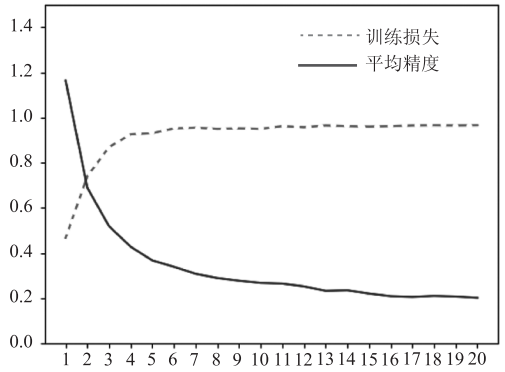
	TOP. 1/%	TOP. 5/%	模型大小/MB	网络深度/层
SqueezeNet1. 0	57. 5	80. 3	0. 72	50
MobileNet V1	70. 6	89. 5	4. 2	28
ShuffleNet1	69. 4	88. 8	5. 4	8
Xception	79	94. 5	88	36
EfficientNet	84. 3	97. 1	66	23

根据对 ImageNet 数据集进行的实验比较，EfficientNet 网络在精度方面表现最优秀。然而，EfficientNet 网络的模型大小仅次于最大的 Xception 网络，相较而言，SqueezeNet1 的模型大小最小，但其网络深度最大，且模型准确度最低。就综合分析识别精度、模型大小、网络深度三个方面而言，MobileNet 和 ShuffleNet1 相对于其他三种轻量化网络表现更优秀，在准确度方面两种相差较小。

为此，对 ShuffleNet 和 MobileNet 模型做进一步实验，采用 Tiny-ImageNet 数据集对两种网络做进一步实验，由于轻量化网络的网络层结构相对简单，采用官方给出的预训练权重，设置训练 20 个 epoch，采用 Python 中的 matplotlib 库对 epoch、accuracy、mean_loss 进行打印，训练结果如图 4 所示，其中横坐标迭代次数，实线表示识别精度曲线，虚线表示训练损失曲线。



(a) ShuffleNet平均精度与训练损失



(b) MobileNet平均精度与训练损失

图 4　ShuffleNet 与 MobileNet 模型训练

Fig. 4 ShuffleNet and MobileNet model training

从图 4 中可以看出，在精度相差较小的情况下，MobileNet 模型的训练损失下降的更快，快速下降的损失通常是模型已经很好地拟合了训练数据，从而导致更好的训练精度和更快的收敛，而 ShuffleNet 的通道重排操作需要额外的计算，这增加了网络的计算量使得网络收敛较慢。因此 MobileNet 更加优秀。

3.2　MobileNet 改进

本文对 MobileNet 进行网络改进，为了进一步应用于红外图像的分类任务，使用自行构建的红外数据集对轻量化网络进行了实验研究。该数据集包括汽车、飞机、摩托车等五个种类的图像，共计 4000 张，其中训练集包含 2500 张，测试集包含 1500 张。在此基础上，以提高 MobileNet 在红外图像分类任务上的性能表现。

在数据格式上，红外图像为单通道，而可见光为 RGB 三通道在卷积层面上输入输出的通道数会不同，并且对于红外图像，细节信息通常是相对较弱的。这是因为红外图像在成像时受到许多因素的影响，如大气湍流、气溶胶、气体吸收和散射等，导致红外图像的噪声和失真较多，特别是在低温场景下。因此，相较于可见光图像，红外图像的细节信息往往比较模糊、不清晰。在这种情况下，如果使用 ReLU6 激活函数，可能会导致一些细节信息被忽略或丢失。而模型需要更好地处理这些细节信息，那么 tanh 函数可能会比 ReLU6 更加适合。本文更改 MobileNet 网络中的激活函数，采用 Tanh 激活函数（如图 5 所示）替换原有的 ReLU6 激活函数。

在更改新的激活函数后，为了能够得到更轻的

网络模型,本文通过减少网络层数的方式减少网络参数。在官方给出的模型中,输入为 224 × 224 × 3 的可见光图像,而在测试应用过程中需要 640 × 512 大小的红外图像输入,为了防止过拟合和增加计算延时的问题,需要重新调整网络参数,使用裁剪网络的方式,更改网络层参数减少网络层结构,成为新的 MobileNet 网络结构(如表 2 所示)。更改后的网络从原来的 28 层缩减为 16 层,增加了卷积核深度,让特征图更快的缩小到指定大小,减少了中间重复的卷积过程,以减少网络模型大小。

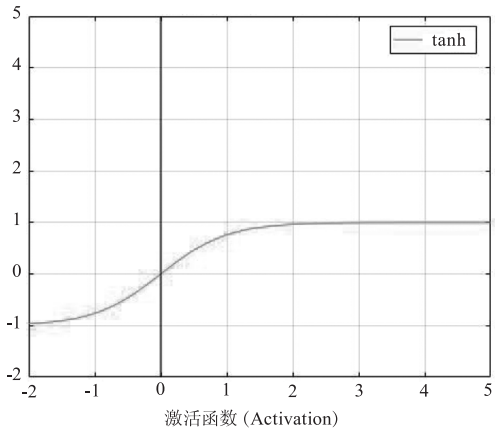


图 5 Tanh 激活函数

Fig. 5 Tanh activation function

表 2 更改后的 MobileNet 网络结构

Tab. 2 The structure of the MobileNet network after modification

Type/Stride	Filter Shape	Input Size
Conv/s2	3 × 3 × 1 × 32	640 × 512 × 1
Conv dw/s2	3 × 3 × 32 dw	320 × 256 × 32
Conv/s1	1 × 1 × 32 × 64	160 × 128 × 64
Conv dw/s2	3 × 3 × 64 dw	160 × 128 × 64
Conv/s1	1 × 1 × 64 × 256	80 × 64 × 64
Conv dw/s2	3 × 3 × 256 dw	80 × 64 × 256
Conv/s1	1 × 1 × 256 × 512	40 × 32 × 256
Conv dw/s1	3 × 3 × 512 dw	40 × 32 × 512
Conv/s1	1 × 1 × 512 × 512	40 × 32 × 512
Conv dw/s2	3 × 3 × 512 dw	40 × 32 × 512
Conv/s1	1 × 1 × 512 × 1024	20 × 16 × 512
Conv dw/s2	3 × 3 × 1024 dw	20 × 16 × 1024
Conv/s1	1 × 1 × 1024 × 1024	10 × 8 × 1024
Avg Pool /s1	Pool 10 × 8	1 × 1 × 1000
FC/s1	1024 × 1000	1 × 1 × 1024
Softmax/s1	Classifier	1 × 1 × 1000

在优化后进行红外图像的测试,测试集包括车、行人、自行车、飞机四种,测试图像先通过对邻近像素的加权平均值来生成新像素来转换同一大小分辨率,测试经 10 个 epoch,最后比较模型准确率、模型大小,实验结果如表 3 所示。

表 3 修改后 MobileNet 表现

Tab. 3 Performance of MobileNet after modification

项目	Train loss	Accuracy	模型大小/MB
MobileNet	0. 354	0. 706	8. 74
修改后 MobileNet	0. 316	0. 651	5. 76

实验对 MobileNet 网络进行了修改,尝试通过改变激活函数和网络层数来提高模型性能。实验结果显示,经过修改的 MobileNet 网络在准确度和模型大小方面都有了一定程度的优化。在原有 MobileNet 网络的基础上,使用 Tanh 作为激活函数,并减少了网络的深度。实验结果显示模型的准确度下降了 0. 052,但模型大小进一步减小了约 3 MB。结果表明,修改激活函数和网络层数可以在一定程度上改善 MobileNet 网络的性能,并且在模型大小和准确度之间需要进行权衡。这表明优化结果具有一定的可行性。接下来利用 FPGA + ARM 结构实现修改后的网络模型,使其能够搭载在硬件上运行。

4 硬件实现

4. 1 实验环境

实验采用 vivado HLS 2018. 3 软件,使用 Xilinx Zynq-7020 XA 开发板进行实验,在 100MHz 工作频率下,实验中设计输入为 640 × 512 的图像输入,其中包含普通卷积、深度可分离卷积、全连接层、池化层最后生成 IP 结合 vivado 软件进行测试。

4. 2 系统架构

采用 ARM + FPGA 的架构,ARM 主要进行初始配置和结果处理,主要运算由 FPGA 完成,ARM 与 FPGA 间采用 AXI 通信协议进行数据传输。卷积模块是整个网络的重点,由于 FPGA 内资源有限,而个层之间运算相似,故设计普通卷积、深度可分离卷积、全连接等进行复用,系统框架图如图 6 所示。

在 PL 部分中设计双端口数据存储,将输入特征 feature map 在读取时拆分为 in1、in2,再在内部定位输入位置进行整合,增加数据的并行性。为了掩盖数据传输时间,采用乒乓操作对数据缓存。将输

入图像的分块按顺序送入计算,分块的大小为 Tp × Tp,每次计算完一个分块的输出后,将输出结果存储到输出缓冲区的对应位置。并且为了防止数据溢出,每次计算完一个分块的输出后,切换缓冲区。以避免了内存空间的浪费,提高了计算效率。

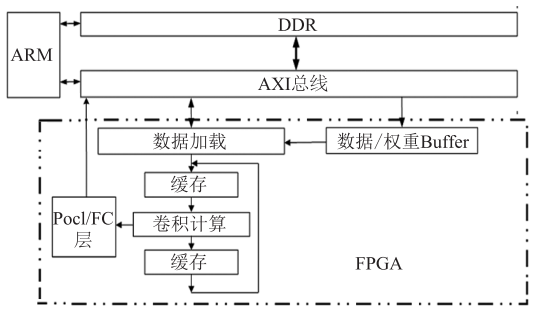


图 6 系统框架图

Fig. 6 System framework diagram

数据的存储与调用采用数据流指令进行调用,以实现个操作直接的并行执行。此外当处理器需要使用一个数据时,它可以从内存中获取该数据并存储在缓存中。如果该数据在后续的计算中仍然需要使用,则不必再次从内存中读取该数据,而是直接从缓存中读取,这样可以显著减少内存访问和数据传输。由于数据量大,网络提高计算效率需要采用数据流水线的方式进行处理,将计算过程分成多个阶段,每个阶段完成一定的计算任务,然后将结果传递给下一个阶段,直到整个计算过程完成,数据流水线如图 9 所示。

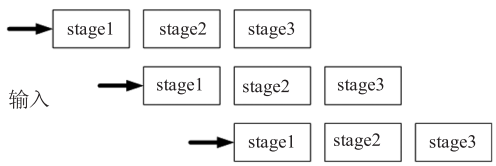


图 7 数据流水线

Fig. 7 Data pipeline

4.3 结果分析

本文最终实现表 2 所示的改进后的 MobileNet 网络,将数据与权重拷贝进 SD 卡中,输入图像为 640 × 512,在 Xilinx Zynq-7020 XA 开发板上 100 MHz 频率下实现每幅图 5.1 ms,经对比在文献[12]中 MobileNet 网络在 100 MHz 工作频率下能实现每幅图像 7.4 ms,本文改进后的 MobileNet 实时性更高,后续还能继续针对算法进行改进,在保障更低延时下实现更高的图像识别准确率。

5 总 结

本文首先通过对轻量化网络的介绍,引入轻量化网络在计算机视觉中的重要性,并对此展开研究,通过轻量化网络在 ImageNet 数据集上的实验,选择出性能最优的轻量化网络——MobileNet 网络,再经过对 MobileNet 网络进行改进,加入 Tanh 激活函数,并降低模型层数,使得能适应红外数据,最后通过 FPGA 对改进后的 MobileNet 进行硬件实现。最后与 MobileNet 相比,改进后的模型精度下降了 0.1,在单张图像大小为 640 × 512 下实现每幅图像 5.1 ms 的计算时间。

针对后续工作,我们认为当前所研究的分类网络的精度还有进一步提高的空间。此外,本文针对该算法在 FPGA 实现 MobileNet 的改进,实际的效果与应用还需要进一步的优化,因此在算法在硬件上的实现与加速方面,仍需要更多的实验和更深入的研究。ShuffleNet 的通道重排技术能够极大的降低网络层结构,减少参数量,但其在 FPGA 上的硬件实现还有一定的困难,后续将在 FPGA 上的实现 ShuffleNet 与 MobileNet 并进行比较,对比两种模型的硬件实现从而进一步优化,最终在硬件上实现一种高实时性的图像识别检测算法。

参考文献:

[1] Jia han. Research on hardware acceleration of lightweight deep convolutional neural networks [D]. Wuhan: Huazhong University of Science and Technology. 2021. (in Chinese)
贾涵. 轻量化深层卷积神经网络的硬件加速技术研究[D]. 武汉:华中科技大学,2021.

[2] LI Cen. HE Guang-hui. Design of FPGA-based neural network accelerator for real-time objective detection[J]. Microelectronics & Computer. 2020,37(7):6 – 11. (in Chinese)
李岑,贺光辉. 用于实时目标检测的 FPGA 神经网络加速器设计[J]. 微电子学与计算机 , 2020,37(7):6 – 11.

[3] Szegedy C, Liu W, Jia Y, Iet al. Going deeper with convolutions [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015:1 – 9.

[4] Iandola F N, Han S, Moskewicz M W, et al. SqueezeNet:

AlexNet-level accuracy with 50 × fewer parameters and <0. 5 MB model size[J/OL]. http://arxiv. org/pdf/1602. 07360.

[5] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arxiv preprint arxiv:1704. 04861, 2017.

[6] Zhang X, Zhou X, Lin M, et al. Shufflenet: an extremely efficient convolutional neural network for mobile devices [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 6848 – 6856.

[7] Chollet F. Xception: Deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1251 – 1258.

[8] Tan M, Le Q. Efficientnet: rethinking model scaling for convolutional neural networks[C]//International Conference on Machine Learning. PMLR, 2019: 6105 – 6114.

[9] ZHOU Guanyu. A high-performance MobileNet accelerator based on FPGA [J]. Machine Building and Automation, 2022, (3): 51. (in Chinese)
周冠宇. 一种基于 FPGA 的高性能 MobileNet 加速器[J]. 机械制造与自动化, 2022, (3): 51.

[10] Lecun Y, Bottou L . Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11): 2278 – 2324.

[11] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770 – 778.

[12] Xiao H, Li K, Zhu M. FPGA-based scalable and highly concurrent convolutional neural network acceleration [C]//2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA). IEEE, 2021: 367 – 370.