轻量化模型在移动端计算机检测系统中的压缩与加速技术

张俊

杭州旭辐检测技术有限公司，浙江杭州，310022；

摘要：随着人工智能技术的广泛应用，计算机视觉模型已在各类终端设备中展现出巨大潜力，尤其是在移动端检测系统中发挥着越来越重要的作用。然而，传统深度神经网络模型因参数量庞大和计算开销高，难以直接部署到资源受限的移动设备上，限制了其实时性与应用场景的拓展。因此，研究轻量化模型及其压缩与加速技术成为当前的研究热点。本文系统地探讨了适用于移动端计算机检测系统的模型压缩方法、推理加速机制与部署优化策略，分别从网络结构重设计、模型剪枝与量化、编译器优化与硬件协同等方面进行详细阐述，旨在为轻量化模型在边缘计算中的实用部署提供理论支持与技术路径。实验表明，经过系统优化的轻量化检测模型在保持较高精度的同时，显著提升了运行效率和资源利用率，具备广泛的工程应用前景。

关键词：轻量化模型；移动端检测；模型压缩；加速推理；边缘计算；神经网络优化

DOI：10.69979/3041-0673.25.10.060

1 轻量化模型结构设计策略

1.1 基于模块化设计的网络简化方法

模块化设计属于一种在轻量神经网络里被广泛运用的结构优化思路，像 MobileNet 系列、ShuffleNet、GhostNet 等都算是其代表性成果。这类模型会把传统的卷积操作分解成更为高效的深度可分离卷积、逐层卷积或者组卷积等不同形式，如此一来便能极大地减少参数的总量以及计算的复杂度。就拿 MobileNet 来说，它运用了深度可分离卷积，把标准卷积分成深度卷积和逐点卷积这两个步骤，进而在精度没有明显降低的情况下，显著地削减了运算量。ShuffleNet 则更进一步，通过引入通道重排操作，也就是 channel shuffle，去改善由组卷积所带来的通道信息被割裂的问题，使得网络中的信息流动能够更加顺畅无阻。在实际进行部署的时候，这类模块化设计策略一方面提升了模型在移动端的运行效率，另一方面也为后续的压缩与加速筑牢了颇为良好的结构基础。

1.2 多尺度特征融合与冗余结构去除

多尺度特征融合在目标检测领域属于提升鲁棒性的关键手段，不过，它也会引发特征冗余以及计算冗余之类的问题。针对这一情况，相关研究人员便提出了一些效率更高的特征融合结构，像轻量化的 FPN（也就是 Feature Pyramid Network）以及动态融合机制等。这些方法在对多尺度信息加以提取的时候，会借助剪枝策略或者注意力机制来有选择性地留存那些有效的特征，以此避免出现重复计算的情况。而且，要是能对卷积核的大小、层数的深度以及通道的维度进行灵活的调整，那么就能够在精度和效率二者之间寻找到更为理想的平衡点。就拿 YOLOv5 的轻量化版本来说，其通过结构剪枝的方式去掉了部分残差连接以及重复的通道，从而将整个模型压缩到了原来体积的 30%，并且仅仅损失了非常微小的精度。这种凭借冗余识别以及结构优化的做法在实际的部署当中所取得的效果是十分显著的。

1.3 基于神经结构搜索的自动化网络裁剪

近些年来，神经结构搜索（也就是 NAS，即 Neural Architecture Search）在诸多场景下被普遍运用起来，其用途主要在于能够自动去发现那些既轻巧又具备较高效率的网络结构。它和传统依靠人工来设计网络的方式是不一样的，NAS 会通过对搜索空间加以明确界定、确定好评估函数以及制定出优化策略等一系列操作，在大规模的搜索活动当中自动去探寻最为优异的方案，进而获取到能够契合目标任务的那种最优架构。在朝着轻量化发展的这个方向上，MobileNetV3 这样的结构就是经由 NAS 和 NetAdapt 相互结合之后才产生出来的，其在性能方面以及效率层面都有着不错的表现。这种方法不但能够在不同的计算预算情形之下生成那种可以按照具体需求定制的模型，而且还能够与不同硬件平台所具有的特性所带来的约束条件很好地相适应，从而使得模型部署时的适配程度得以有效提升。伴随着搜索效率持续地被优化改进，NAS 正在一步步地变成轻量模型设计领域当中主流的方法其中之一，也为移动端模型的开发开拓出了一些新的可能途径。

2 模型压缩与推理加速方法

2.1 权重量化与激活离散化技术

权重量化属于轻量化模型当中极为关键的一项核心技术。它主要是要把高精度的浮点权重映射成低比特定点值，像 INT8 或者 INT4 这类的，以此来让模型的存储需求得以降低，同时也能使计算复杂度有所下降。对网络参数展开线性或者非线性的量化操作，是能够切实有效地减少内存带宽方面的消耗情况的，而且还可以在一定程度上提升运算的效率，另外还能兼顾到让模型精度维持在一个较为合适的水平。就比如说，TensorRT 以及 ONNX Runtime 等这些框架，它们就给出了高效的量化工具链，能够支持针对训练后模型去做离线量化，或者是结合校准数据来开展量化感知训练，也就是常说的 QAT。

除此之外，把激活值进行离散化处理，这同样也是提高推理速度的一个重要策略。通过将激活范围缩小，并且降低其精度，也能够比较明显地对中间特征图所占用的存储空间进行压缩。在进行量化的这个过程当中，为了能够减轻精度方面可能出现的损失情况，常常会把剪枝以及蒸馏等相关技术结合起来协同开展优化工作，从而保证模型在移动端进行推理的时候，既能够实现快速运算，又可以确保其准确性。

2.2 稀疏剪枝与通道重构策略

模型剪枝技术能够识别并去掉网络当中那些冗余的参数或者连接结构，以此达成对模型规模的明显压缩效果。在轻量化模型这块儿，剪枝大体上可以分成权重剪枝、通道剪枝以及结构剪枝这三类情况。权重剪枝主要是针对单个连接来开展相关操作的，通道剪枝呢则是面向整个卷积通道去进行处理，至于结构剪枝它是可以跳过整层运算的。稀疏剪枝会借助像 L1 正则化或者梯度掩码等这类技术，促使网络自然而然地朝着稀疏化的方向发展，随后在推理阶段就能够把那些值为零的参数彻底给移除掉，这样一来就能加快模型执行的速度。在通道剪枝方面，通过对通道重要性进行评估，比如说参考 BN 缩放系数、梯度信息等等这些内容，就可以从中筛选出关键的路径，从而能够较为有效地防止出现信息丢失的情况。而且，剪枝之后的网络往往还得借助通道重构策略来再次开展训练以及进行微调方面的工作，目的就是为了恢复或者提升其精度。这些技术为轻量化模型在移动端能够快速地进行部署提供了挺实用的办法。

2.3 蒸馏压缩与知识迁移方法

知识蒸馏属于那种以教师学生模型当作框架的模型压缩形式，其是借助训练小模型去模仿大模型的输出表现，进而达成对知识予以有效迁移的目的。该方法一方面能够提升小模型的性能，另一方面还拥有比较不错的泛化能力。常见的蒸馏技术包含软标签蒸馏、特征层对齐以及注意力迁移等不同方式。就软标签蒸馏来讲，它是通过将学生模型与教师模型在输出层的概率分布差异尽量缩小，促使学生模型能够更贴近真实数据的分布状态。而特征层对齐则是进一步推动蒸馏过程往深处发展，让中间层特征具备更为突出的表达能力。当知识蒸馏和剪枝、量化搭配起来使用的时候，它可以充当精度恢复的重要环节，尤其在边缘设备部署之前的模型压缩优化阶段，它是格外适用的。因其具备简单且高效的特点，同时还易于集成，所以在当下的多个实际工业项目当中，已经得到了十分广泛的应用。

3 移动端部署与加速框架优化

3.1 边缘设备异构特性下的优化策略

移动端设备在计算资源这块呈现出的多样性，着实为轻量化模型的部署赋予了相当高的灵活性，与此同时，也相应地带来了适配方面更高程度的挑战。像现代智能手机、可穿戴设备以及嵌入式终端等，一般都会配备诸如 CPU、GPU、NPU 或者 DSP 等多种不同的计算核心。这些硬件单元，各自有着自身在处理方面的优势，当然了，也存在着一定的局限性。所以在进行模型部署的时候，务必要依据设备实际所具有的配置情况来开展差异化的优化操作。就拿 Android 平台来举例说明，TensorFlow Lite 和 NNAPI 接口相互结合之后，是能够对系统当中的 AI 加速器加以识别的，并且还可以自动地把模型推理任务调度到最为优质的硬件资源上去，如此一来，便能够在很大程度上提升计算的效率。而要是说到 iOS 平台的话，CoreML 通过深度整合 Metal GPU 架构以及 Apple Neural Engine，是可以达成一种具备低延迟、低功耗特点的神经网络执行机制的。

3.2 基于图优化与编译器加速的部署工具链

在模型训练完毕且做完结构压缩之后，要是没有有效的编译以及部署工具链给予支持的话，那么它的推理性能就很有可能会因为操作没能融合起来、内存使用不太合理或者算子调用太过频繁这些情况，进而受到极为严重的制约。图优化和深度学习编译器出现的主要目的就是要来解决这一难题，其关键之处就在于对模型的计算图去做结构方面的重新构建、把算子进行融合处理、对数据调度加以优化以及开展指令级的并行处理等一

系列操作，通过这样的方式来促使模型在实际的设备上运行效率得以提升。就拿 TVM 当作代表的开源编译器来说，它是借助 Relay IR 来表示模型的结构，在编译的整个过程当中，会结合目标硬件所具有的特征去展开自动调度策略的搜索行动，如此一来，便能够给不同的芯片生成最为理想的低层代码。而 TensorRT，它相对来说是更侧重于 NVIDIA GPU 平台的，通过采取子图融合、对权重进行精简、转换数据格式等诸多机制，把高精度的深度模型转变成适配性特别高的高性能执行图，这样就能在很大程度上提升推理的速度，并且还能够让延迟得以减少。

在同一时期，XLA 以及 OpenVINO 这类编译器，它们分别针对 Google TPU 以及 Intel 架构展开了专门的图级优化操作，进而构建起与模型结构、数据类型还有运行时硬件都紧密关联起来的加速体系。这些工具链一方面使得开发者在部署之时，无需再去进行那些关于底层实现的繁杂操作，另一方面也在很大程度上降低了模型从实验室环境过渡到产品级应用场景的迁移难度，它们实实在在地成为了达成轻量化模型高效部署的关键技术支撑所在。

3.3 多任务融合与端侧协同推理机制

随着移动设备智能化程度持续提高，在其实际应用当中常常得同时开展多个计算机视觉方面的任务，像是图像识别、物体检测、姿态估计，甚至还包括多模态信息处理等。为了能契合这种多任务的实际需求，轻量化模型正逐步朝着共享主干网络以及分支多头结构这个方向去发展演变，具体来讲，就是依据一套统一的特征提取网络，针对不同的任务去精心设计高效的解码头部，如此一来，多个任务便能够同时并行处理，进而节省下那些冗余的计算资源。这种融合式的设计不光提高了整体模型的参数复用率以及存储效率，而且还在相应程度上强化了任务之间的协同能力与场景适应能力。

在实际的系统当中，为了能够更好地兼顾实时性以及精度这两方面的要求，端云协同推理机制已然慢慢变成了主流的方案。具体来说，就是把那些计算复杂度比较高，而且对延迟不太敏感的任务转移到云端去加以处理，而在移动端本地，只去执行那些对延迟敏感或者和隐私相关的核心功能。通过这样的方式，就可以有效地去分担计算方面的负荷，还能降低设备的能耗，并且也有助于提升用户的交互体验。该机制一般会结合轻量化模型的分段部署策略以及特征中间层缓存的做法，来让部分中间结果在端侧得以生成，随后经过优化编码之后发送到云端，以便完成后续的推理工作，最后再把结果返回到本地进行展示。这种端云协同再加上多任务融合的架构，在很大程度上扩展了轻量模型在复杂场景之下的适用范围。特别是在智能交通、工业巡检、远程医疗以及增强现实等诸多场景之中，它都展现出了极为优异的系统性能以及资源平衡方面的能力，是推动下一代移动智能系统不断演进的一条关键的技术路径。

4 结语

随着人工智能应用不断向边缘侧延伸，轻量化模型已成为推动移动端计算机视觉系统发展的核心力量。本文从模型结构设计、压缩与加速技术、部署优化策略三个方面系统探讨了轻量化模型在移动端检测系统中的关键技术路径与实践方法。通过对模块化网络、神经结构搜索、量化剪枝、知识蒸馏、编译器优化以及异构硬件适配等策略的综合应用，当前的轻量化模型不仅在精度上接近甚至超越传统大型模型，更在推理速度、资源占用与能耗方面取得显著优化，具备广泛的实际部署价值。

然而，轻量化模型的发展仍面临一些挑战，如压缩过程中精度损失的平衡、自动化搜索的效率与可靠性、异构设备兼容性问题等。未来的研究可以从以下几个方向展开：一是探索更具通用性与鲁棒性的结构搜索算法，实现跨任务轻量模型自适应构建；二是进一步强化边云协同推理机制，优化系统整体延迟与资源分配；三是结合大模型知识迁移与小模型部署需求，发展更加高效的知识蒸馏框架。总的来看，随着软硬件技术的持续进步，轻量化模型将在智能终端的视觉感知与决策中扮演越来越重要的角色，推动移动人工智能技术进入高性能、低功耗、多任务融合的新阶段。

参考文献

[1]赵家祥.企业 ESG 评级对股票市场资产定价偏误的影响研究——基于中国 A 股上市公司[J].商业观察,2025,11(08):104-110+115.
[2]孙如雪.企业 ESG 表现与资产误定价[J].商业观察,2025,11(05):83-88.
[3]乔宏.考虑客户忠诚度与企业成本削减的产品定价策略[J].全国流通经济,2025,(03):73-76.
[4]王珍珍.探究消费者耐心程度与企业成本削减对产品定价策略的影响[J].全国流通经济,2025,(03):81-84.