ISSN：3078-8994
EISSN：3104-4492

轻量化神经网络图像识别简易方案研究

李 淳

山西财经大学信息学院，山西 太原 030000

摘 要：传统卷积神经网络往往参数规模较大、计算负担也偏高，因此在移动终端与嵌入式设备等资源受限的场景中不易落地应用，促使图像识别模型从重载复杂逐步转向轻量高效，已经成为计算机视觉领域较为重要的研究方向。本文依照规范的学术论文写作框架，对轻量化神经网络的核心设计思路与主要实现技术进行较为系统的说明，并在此基础上搭建轻量化图像识别方案的整体框架，以深度可分离卷积作为核心手段、以模型压缩作为辅助方式、以简易部署作为目标导向，围绕网络结构设计、模型压缩与优化、推理加速与部署等关键实现路径展开研究，进一步提出面向资源受限场景的简易方案技术体系，期望为移动端与嵌入式设备中的图像识别应用提供低成本、高效率的可行方案与实践参考。

关键词：轻量化神经网络；图像识别；深度可分离卷积；模型压缩；边缘部署

DOI：10.64649/yh.shfzykjcx.issn3078-8994.202606017

1 研究背景与意义

近年来，深度学习在图像识别方向上取得比较突出的成果，以 ResNet、VGG 等为代表的传统卷积神经网络在大型数据集上的识别精度表现很强，但这类模型往往带有数千万甚至上亿的参数，同时需要 GB 级内存以及数十亿次浮点运算，这就对其在智能手机、物联网节点等资源受限的环境中部署造成了明显限制。随着边缘计算的快速推进，图像识别在实时性、高效率、低功耗等方面的需求变得更加紧迫，于是轻量化神经网络逐渐出现并被关注，轻量化网络通常通过结构上的改动与模型压缩等方式，在保持可接受精度的条件下明显减少参数量与计算量，使深度学习模型不再只停留在服务器机房中，而是能够进入移动终端与各类边缘设备。开展本研究在一定程度上可以降低技术应用的门槛，并推动计算机视觉技术在智能安防、工业质检、智慧农业等场景中实现更大范围的落地。

2 轻量化神经网络现状与转型需求

2.1 传统神经网络的应用困境

传统卷积神经网络主要依靠把卷积层与全连接层不断加深、加厚来构建模型，通过提升网络的深度和宽度来增强特征表达能力。以 VGG16 为例，其参数量超过 1.38 亿，模型体积在 500MB 以上，单次前向推理大约需要 155 亿次浮点运算。在 GPU 服务器上这类模型通常可以较高效地运行，但放到资源受限的平台时，会遇到存储空间不够、推理速度偏慢、功耗偏高这三重困境：第一，模型体积过大，往往难以烧录进设备的闪存；第二，计算量过大，推理延迟容易超过实时性的要求；第三，高功耗的运行状态会明显缩短移动设备的续航时间。由此可见，上述不足使传统神经网络较难满足智能终端在图像识别方面对实时性、轻量化以及低功耗的需求。

2.2 轻量化神经网络的内涵与核心优势

轻量化神经网络主要强调结构更简、计算更省、部署更方便，它依靠新型的网络架构设计和模型压缩等方法，在资源比较有限的情况下也能做到基本可用的识别效果。其优势大致表现在：首先，参数数量明显减少，使得模型体积能够压到数 MB 级别，从而更容易适配移动端的存储限制；其次，计算复杂度被进一步压低，推理速度可以满足实时处理的要求；第三，功耗相对可控，更适合需要长时间持续运行的边缘应用场景。总体来看，轻量化网络让深度学习模型在算力、存储与功耗都受限的环境里仍能保持有效运行，也为计算机视觉技术更大范围的推广应用提供了一定的基础支撑。

3 轻量化图像识别方案总体架构

3.1 轻量化方案的核心逻辑与技术路线

轻量化图像识别的简易方案一般遵循“结构轻量化—训练轻量化—部署轻量化”这三阶段的技术路线，首先在结构轻量化方面，以深度可分离卷积、分组卷积、倒残差结构等作为主要组件，从一开始就尽量降低模型的参数量与计算量，其次在训练轻量化方面，通过知识蒸馏、迁移学习等方法，让大模型对小模型起到指导学习的作用，从而提高轻量模型的泛化能力，最后在部署轻量化环节，借助模型量化、剪枝、算子融合等技术，进一步压缩模型体积并提升推理速度，这三个阶段前后衔接，逐层递进，也呈现出逐层递进的关系，共同构成该

方案的技术逻辑主线。

3.2 轻量化方案的评价指标体系

轻量化图像识别方案一般需要建立一个多维度的评价体系：在精度方面，常用指标包括 Top-1 准确率、Top-5 准确率；在效率方面，指标主要有模型参数量（Params）、浮点运算量（FLOPs）、模型体积；在速度方面，通常关注单帧推理延迟、每秒处理帧数（FPS）；在部署方面，则会看内存占用、功耗水平、框架兼容性。以上四类指标在不同应用场景中需要先确定各自的优先级，再做相应的权衡策略，例如工业质检更强调精度与速度之间的均衡，安防监控更关注实时性与稳定性，移动端应用则更重视体积与功耗。

3.3 四层一体总体架构设计

轻量化图像识别的简易方案大体采用四层一体的架构：数据层主要承担图像数据的获取、预处理和增强工作，包括尺寸归一化与数据扩增，以便提升模型的泛化能力；模型层整合 MobileNet、ShuffleNet 等常见的轻量网络，允许用户根据需求进行选择与参数配置；优化层提供模型剪枝、量化、蒸馏等压缩手段，用来实现模型的进一步轻量化；部署层则连接 TensorFlow Lite、ONNX Runtime、NCNN 等推理引擎，支持在多平台进行输出与落地。整体来看，这四层架构在纵向上能够贯通流程，在横向上又相对解耦，从数据准备一直到终端部署形成一套较为完整的一站式简易方案，从而降低开发者的技术门槛与实施复杂度。

4 轻量化图像识别关键技术方法

4.1 深度可分离卷积与轻量网络设计

深度可分离卷积是轻量化神经网络的核心技术，将标准卷积分解为深度卷积与逐点卷积两步。标准卷积的参数量为 $D_K \times D_K \times C_{in} \times C_{out}$，深度可分离卷积的参数量为 $D_K \times D_K \times C_{in} + C_{in} \times C_{out}$，计算量压缩至原来的约 $1/C_{out} + 1/D^2_K$。MobileNetV 利用该结构将标准卷积 FLOPs 从 558 M 降至 56 M 。 MobileNetV2 在此基础上引入倒残差结构（Inverted Residual Block），采用 " 扩张一卷积一压缩 " 的沙漏结构：先用 1×卷积扩展通道数，再进行深度卷积提取特征，最后用 1×1 卷积压缩通道，在保持计算效率的同时提升了特征表达能力。ShuffleNet 通过分组卷积与通道混洗，在超低算力设备上实现了良好精度—效率权衡。

4.2 模型压缩：剪枝、量化与知识蒸馏

模型压缩被认为是继续把模型做得更轻的重要技术。剪枝主要是把多余的神经元、卷积核或者通道移除，从而降低参数量，其中结构化剪枝一般以整个通道或滤波器作为裁剪单位，这样能让模型结构保持相对规整，方便后续做硬件部署，同时也可以将 MobileNetV2 的参数量减少 50%，并把精度损失控制在 1.1% 以内。量化则是把模型权重由 32 位浮点数转换成 8 位整数，使模型体积大约缩小 4 倍，并让推理速度提升 2 至 3 倍，量化感知训练是在训练阶段模拟量化误差，从而在一定程度上减少精度损失。知识蒸馏是让参数量更大、精度更高的教师网络去指导轻量的学生网络学习，把教师网络中的暗知识迁移到学生网络中，这样在保持体积紧凑的同时，也能较明显地提升识别精度。

4.3 简易部署与推理加速方案

轻量化模型的最后目的在于更方便地完成部署，目前移动端常见的推理框架主要有 TensorFlow Lite、ONNX Runtime、NCNN、MNN 等，这些框架一般都支持模型格式的转换，也能配合硬件加速来提升推理效率，部署阶段的优化做法包括：首先是算子融合，把卷积、批归一化、激活函数等环节合并成一个算子，从而减少计算图里的节点数量，并降低内存访问带来的开销，其次是内存复用，通过张量的内存池化来减少动态分配的负担，第三是多线程并行，借助 ARM NEON 等指令集来加速卷积相关的计算，在一些资源极端受限的应用场景中，还可以采用更极简的网络结构和更极致的量化方式，把模型压缩到百 KB 这个量级，同时配合特征图像预处理，让输入数据量减少约 97%，并使模型的有效尺寸减少约 98%，分类精度的损失尽量控制在 1.5% 以内。

5 简易方案实施路径与对比分析

5.1 分阶段实施路径

轻量化图像识别简易方案的实施一般分为四个阶段。第一阶段是需求分析与方案选型，需要先把应用场景里的资源限制说清楚，包括算力、内存、功耗，同时也要明确精度要求和实时性指标，再依据这些条件去选择更匹配的轻量网络基座。第二阶段是模型训练与调优，主要是借助迁移学习在目标数据集上对预训练模型做微调，如果有需要也可以引入知识蒸馏来提高精度。第三阶段是模型压缩与优化，要结合部署硬件的特点来决定压缩方式，可以选择剪枝、量化，或者把两者结合起来完成模型压缩。第四阶段是部署测试与迭代，将已经优化过的模型转换成目标平台所需要的格式，然后开展端侧测试与效果评估，再根据反馈继续做迭代优化。按阶段推进有助于降低实施过程中的风险，也更便于尽快验证效果，从而让方

案相对顺利地落地。

5.2 主流轻量网络方案对比

为了便于方案的选择，表 1 对目前较常见的主流轻量神经网络做了较为系统的对照分析。

表1 主流轻量神经网络方案对比

模型	核心创新	参数量	计算量 (FLOPs)	Top-1 准确率	适用场景
MobileNetV1	深度可分离卷积	4.2M	569M	70.60%	通用移动端
MobileNetV2	倒残差结构 + 线性瓶颈	3.5M	300M	72.00%	移动端 / 嵌入式
MobileNetV3	神经架构搜索 +SE 模块	5.4M	219M	75.20%	高性能移动端
ShuffleNetV2	通道混洗 + 分组卷积	2.3M	149M	69.40%	超低算力设备
EfficientNet-B0	复合缩放 + 神经架构搜索	5.3M	390M	77.10%	精度优先场景
SqueezeNet	Fire 模块	1.2M	-	57.50%	极致轻量场景

注：以上数据基于 ImageNet 数据集，不同实现版本与输入尺寸下数值可能存在差异，参数量以 M（百万）为单位。

5.3 简易方案的工具链保障

方案要真正落到实处，需要有一套配套而且相对完整的工具链来支撑。训练阶段可以使用 TensorFlow、PyTorch 等框架中的预训练模型库和迁移学习接口，从而较快完成模型的适配与微调。压缩阶段则能够借助 TensorFlow Model Optimization Toolkit、NNI、PaddleSlim 等开源工具来实现自动化剪枝与量化，手动调参的难度也会随之降低。部署阶段可以选择各类推理引擎所提供的模型转换工具与性能分析器，用来较快完成端侧适配与性能调优。通过推动工具链的标准化与自动化，普通开发者不必深入到底层实现，也可以完成轻量化图像识别方案的构建与部署，从而在一定程度上降低技术应用门槛。

6 结论

本文围绕传统卷积神经网络在资源受限条件下不易落地的问题，提出一套以深度可分离卷积为主要手段、以模型压缩作为补充、并以便于部署为导向的轻量化图像识别方案。研究结果显示，深度可分离卷积能够把计算量降低约 8~9 倍，结构化剪枝可以让参数量减少 50%，量化则使模型体积缩小 4 倍，同时推理速度提升 2~3 倍，三种方法配合使用时，在精度损失相对可控的情况下，可以基本满足移动端部署的需要。进一步而言，文中搭建的四层一体架构以及分阶段的实施路径，为开发者提供了一个技术较成熟、工具链也较完整的参考方案，便于按步骤完成部署工作。

参考文献：

[1] 赵玉亮 . 轻量化神经网络模型在移动设备图像识别上的应用探索 [J]. 中国信息化 ,2025,(10):66-67.

[2] 程玲 , 聂罗娜 . 基于深度神经网络的智能图像识别模型研究进展与优化趋势 [J]. 中国宽带 ,2025,21(7):133-135.

[3] 聂刚刚 , 饶洪辉 , 康丽春 , 等 . 基于轻量化卷积神经网络的油茶病害识别 [J]. 江西农业大学学报 ,2024,46(2):502-515.

[4] 闫涵 , 卢伟 , 吴玉虎 . 基于轻量化卷积神经网络的金属断口图像识别 [J]. 控制与决策 ,2024,39(9):2858-2866.

[5] 熊鹏文 , 陈志远 , 廖俊杰 , 等 . 基于改进卷积注意力机制的触觉图像识别 [J]. 东南大学学报 (自然科学版),2024,54(1):175-182.

作者简介：李淳 (2004.10—)，男，汉族，福建省龙岩市，本科，研究方向：计算机科学与技术。