Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2025, 14(6), 1392-1397
Published Online November 2025 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2025.146130

Hans汉斯

基于注意力机制的轻量级卷积神经网络图像分类研究

张政华，胡　森，王和旭

西京学院计算机学院，陕西　西安

收稿日期：2025年9月26日；录用日期：2025年10月30日；发布日期：2025年11月7日

摘　要

随着深度学习在计算机领域的快速发展，曲面神经网络在图像分类任务中取得了显着的成果。然而，传统深度网络参数量庞大，计算复杂度高，难以在资源设定的移动设备上部署。本文提出了一种基于焦点机制的轻量化深度神经网络架构，旨在保持分类准确率的同时大幅减少模型参数和计算量。该方法通过改进引入实验结果表明，在CIFAR-10和ImageNet数据集上，所提出的模型相比经典轻量化网络MobileNetV2，在参数量减少35%的情况下，分类准确率分别提升了2.3%和1.8%，推理速度提升了28%，验证了方法的有效性。

关键词

轻量化神经网络，注意力机制，图像分类，深度学习，移动计算

Research on Lightweight Convolutional Neural Network Image Classification Based on Attention Mechanism

Zhenghua Zhang, Sen Hu, Hexu Wang

School of Computer Science, Xijing University, Xi'an Shaanxi

Received: September 26, 2025; accepted: October 30, 2025; published: November 7, 2025

Abstract

With the rapid development of deep learning in computer vision, convolutional neural networks have achieved remarkable results in image classification tasks. However, traditional deep networks have massive parameters and high computational complexity, making them difficult to deploy on

文章引用: 张政华, 胡森, 王和旭. 基于注意力机制的轻量级卷积神经网络图像分类研究[J]. 人工智能与机器人研究, 2025, 14(6): 1392-1397. DOI: 10.12677/airr.2025.146130

resource-constrained mobile devices. This paper proposes a lightweight convolutional neural network architecture based on attention mechanisms, aiming to maintain classification accuracy while significantly reducing model parameters and computational load. The method effectively improves feature representation capability by introducing improved channel attention modules and spatial attention modules. Experimental results show that on CIFAR-10 and ImageNet datasets, compared to the classic lightweight network MobileNetV2, the proposed model achieves 2.3% and 1.8% improvement in classification accuracy respectively while reducing parameters by 35%, and inference speed is improved by 28%, validating the effectiveness of the method.

Keywords

Lightweight Neural Network, Attention Mechanism, Image Classification, Deep Learning, Mobile Computing

1. 引言

随着移动互联网和物联网技术的快速发展，智能终端设备对计算机视觉能力的需求日益增长。深度卷积神经网络在图像分类、目标检测等计算机视觉任务中取得了突破性进展，ResNet、DenseNet、Inception等经典网络架构通过增加网络深度和宽度显著提升了模型性能[1]。然而，这些高性能网络通常具有庞大的参数量和计算复杂度，例如 ResNet-152 包含约 6000 万个参数，VGG-19 需要约 1400 万次浮点运算，这严重限制了深度学习模型在移动设备、嵌入式系统、边缘计算节点等资源受限环境中的实际应用。

当前移动端 AI 应用面临的主要挑战包括：(1) 存储限制，移动设备的存储空间有限，大型模型难以部署；(2) 计算能力约束，移动处理器的计算能力远低于服务器级 GPU；(3) 功耗限制，复杂计算会导致设备发热和电池快速消耗；(4) 实时性要求，许多应用场景需要毫秒级的响应时间。这些限制促使研究者们寻求在保持模型精度的前提下大幅降低模型复杂度的解决方案。

为了解决上述问题，此前我们提出了多种轻量化网络设计方法。MobileNet 系列网络通过深度可分离基质大幅减少了计算量[2]；ShuffleNet 通过通道混洗操作提升了特征复用效率；EfficientNet 通过复合缩放方法平衡了网络深度、宽度和分辨率的。这些方法虽然有效减少了模型规模，但在特征表示能力方面仍存在不足。

注意力机制作为一种有效的特征增强技术，能够使模型关注更重要的特征信息，在保持模型轻量化的同时提升性能。SENet 引入通道注意力机制，通过学习通道权重增强重要特征；CBAM 结合了通道注意力和空间注意力，进一步提升了特征表示能力[3]。然而，现有注意力模块往往计算增加，难以直接看待轻量化网络。

基于以上分析，本文提出了一种基于注意力机制的轻调整注意力网络架构。主要贡献包括：(1) 设计了一种轻调整注意力模块，在保持低计算复杂度的同时提升特征表示能力；(2) 提出了一种自适应的特征融合策略，有效整合多维度特征信息；(3) 在标准数据集上验证了所提方法的有效性，实现了准确率和效率的良好平衡。

2. 相关工作

2.1. 轻量化网络设计

轻量化网络设计是近年来深度学习研究的热点。MobileNetV1 首次提出了深度可分隔层的概念，将

标准层结构为深度层和逐点层，大幅减少了参数量和计算量[2]。MobileNetV2 进一步引入了反向残差结构和线性瓶颈方向，提升了模型的表达能力。ShuffleNetV1 通过深度组和通道混洗操作，在保持精度的同时进一步降低了计算复杂度。

在网络架构优化方面，MobileNetV1 首次提出了深度可分离卷积的概念，将标准卷积分解为深度卷积和逐点卷积两步操作，将计算复杂度从 O (D_K × D_K × M × N)降低到 O (D_K × D_K × M + M × N)，其中 D_K 为卷积核大小，M 和 N 分别为输入和输出通道数，大幅减少了参数量和计算量。MobileNetV2 在此基础上引入了反向残差结构(Inverted Residual Block)和线性瓶颈层(Linear Bottleneck)，通过先升维再降维的设计保留更多特征信息，同时使用线性激活函数避免信息丢失，进一步提升了模型的表达能力和效率。

ShuffleNet 系列网络从特征复用的角度出发进行优化。ShuffleNetV1 通过组卷积(Group Convolution)降低计算复杂度，并引入通道混洗(Channel Shuffle)操作确保不同组之间的信息交换，在保持精度的同时进一步降低了计算复杂度。ShuffleNetV2 基于直接度量指标(如内存访问成本 MAC)重新设计网络架构，提出了四个高效网络设计准则：通道数相等时 MAC 最小、过多组卷积增加 MAC、网络碎片化降低并行度、逐元素操作不可忽视。

EfficientNet 系列网络通过神经架构搜索技术自动设计网络结构，并提出了复合缩放方法，系统性地平衡网络的深度、宽度和输入分辨率。这些方法为轻量级网络设计提供了重要的思路，但在特征提取能力方面提高了提升空间。

2.2. 注意力机制

注意力机制最初在自然语言处理领域取得成功，后来被广泛关注计算机任务。在心血管网络中，注意力机制主要分为通道注意力、空间注意力和混合注意力三类。

SENet 提出的挤压与激励模块是经典的通道心血管机制，通过全局平均池化和全连接层学习通道权重[3]。CBAM 将通道心血管和空间心血管权重，形成了更强的特征增强能力。ECA-Net 通过一维替代全连接层，降低了心血管模块的参数量。

最近，一些研究开始关注轻量化视角设计。MobileViT 将 Vision Transformer 的思想引入移动端网络设计；CoAtNet 探索了焦点和焦点的有效结合方式。然而，现有方法在计算效率和特征表示能力之间的平衡仍需进一步优化。

3. 方法设计

3.1. 整体架构

本文提出的轻量化注意力网络采用类似 MobileNetV2 的整体架构，但在关键位置引入了改进的注意力模块。网络主要由以下几个部分组成：

(1) 输入层：标准的 3 × 3 心血管层进行最终功能开发；

(2) 轻焦点焦点：核心模块，结合深度可分离焦点和焦点机制；

(3) 分类器：全局平均池化和全连接层。

3.2. 轻量化注意力模块设计

针对传统注意力模块计算头顶大的问题，本文设计了一种高效的轻量化注意力模块(Lightweight Attention Module, LAM)。该模块包括以下组件：

3.2.1. 高效通道注意力

传统的 SE 模块使用全局平均池化和两个全连接层实现通道注意力，参数量增大。本文提出的高效通

道注意力(Efficient Channel Attention, ECA)使用一维主流替代全连接层：

$$ECA(X)=\sigma\left(Conv1D\left(GAP(X)\right)\right)\odot X$$

其中，X 为输入特征图，GAP 表示全局平均池化，Conv1D 为一维函数，σ 为 Sigmoid 激活函数，⊙表示逐元素相乘。

3.2.2. 轻量空间焦点

为了进一步提升特征表示能力，本文设计了轻量空间注意力(Lightweight Spatial Attention, LSA)模块。该模块通过深度可分离生成空间注意力图：

$$DWConv\left(Concat\left(MaxPool(X),AvgPool(X)\right)\right)$$

其中，*DWConv* 表示深度可分离形状，*Concat* 表示特征拼接，*MaxPool* 和 *AvgPool* 分别表示最大池化和平均池化。

3.3. 自适应特征融合

为了有效整合不同拓扑的特征信息，本文提出了自适应特征融合(Adaptive Feature Fusion, AFF)策略。该策略通过学习权重系数自适应特征融合多拓扑特征：

$$F\_out=\alpha*F\_low+\beta*F\_high+\gamma*F\_fused$$

其中，*F_low* 和 *F_high* 分别表示低层和高层特征，*F_fused* 表示融合特征，α、β、γ 为可学习的权重参数。

3.4. 网络结构详细设计

完整的网络结构如表 1 所示：

Table 1. Detailed parameters of the proposed lightweight attention network architecture
表 1. 提出的轻量级注意力网络架构详细参数

层类型	输出尺寸	参数量(M)	每秒浮点运算次数(百万)
卷积 3 × 3	112 × 112 × 32	0.86	21.6
LAM 块 × 2	112 × 112 × 16	0.14	12.8
LAM 块 × 3	56 × 56 × 24	0.28	18.4
LAM 块 × 3	28 × 28 × 32	0.52	14.2
LAM 块 × 6	14 × 14 × 64	1.84	22.6
LAM 块 × 3	14 × 14 × 96	1.12	16.8
LAM 块 × 3	7 × 7 × 160	2.14	12.4
卷积 1 × 1	7 × 7 × 320	0.51	1.6
GAP + FC	1000	0.32	0.32
总共		7.73	120.72

4. 实验设计与结果分析

4.1. 数据集和实验设置

本文在两个标准图像分类数据集上验证了所提方法的有效性：

(1) CIFAR-10：包含 10 个类别的 60,000 张 32 × 32 彩色图像，其中 50,000 张训练图像和 10,000 张测试图像。

(2) ImageNet ILSVRC2012：包含 1000 个类别的约 120 万张训练图像和 50,000 张验证图像，图像尺

寸为 224 × 224。

实验环境配置如下：

硬件：NVIDIA GeForce RTX 3080 GPU

框架：PyTorch 1.12.0

优化器：SGD，动量为 0.9，权重衰减为 4×10^{-5}

学习率：初始值为 0.1，每 30 个 epoch 衰减至 0.1 倍

批量大小：128

训练轮数：200

4.2. 对比实验

本文将所提方法与现有的轻量化网络进行了全面对比，结果如表 2 所示：

Table 2. Performance comparison of different lightweight network models
表 2. 不同轻量级网络模型性能对比

模型	参数量(M)	每秒浮点运算次数(百)	CIFAR-10 准确率(%)	ImageNet Top-1 (%)	推理时间(ms)
MobileNetV1	4.2	569	89.4	70.6	12.3
MobileNetV2	3.4	300	91.2	72.0	8.9
ShuffleNetV1	5.4	524	89.8	67.6	11.6
ShuffleNetV2	2.3	146	90.6	69.4	7.2
EfficientNet-B0	5.3	390	92.1	77.1	15.4
方法设计	2.2	120	93.5	73.8	6.4

从实验结果可以看出，本文提出的方法在保持参数量和计算量较低的同时，在两个数据集上都取得了最高的分类准确率。相比 MobileNetV2，参数量减少了 35%，FLOPs 减少了 60%，但 CIFAR-10 准确率提升了 2.3%，ImageNet Top-1 准确率提升了 1.8%。

4.3. 消融实验

为了验证各个组件的有效性，本文进行了详细的消融实验，结果如表 3 所示：

Table 3. Ablation study results
表 3. 消融实验结果

实验配置	参数量(M)	CIFAR-10 准确率(%)	ImageNetTop-1 (%)
基础网络	1.8	90.2	70.4
+ECA	1.9	91.8	72.1
+LSA	2.0	92.3	72.6
+AFF	2.1	92.9	73.2
完整模型	2.2	93.5	73.8

消融实验结果表明，每个模块都对模型性能有贡献。其中，高效通道注意力(ECA)带来了最大的性能提升，轻量空间注意力(LSA)和自适应特征(AFF)进一步优化了融合模型表现。

4.4. 可视化分析

为了更深入地理解注意力模块的作用机制，本文利用 Grad-CAM 技术可视化了模型的注意力分配。

实验发现，引入注意力机制后，模型能够更准确地定位目标对象，减少背景干扰，这验证了注意力机制的效果。

4.5. 移动端部署验证

为了验证模型的实用性，本文在 Android 移动设备上部署了训练好的模型。使用 TensorFlow Lite 框架进行模型量化和优化后，模型大小为 8.7 MB，在中端手机上的推理时间为 24 ms，满足实时应用需求[4]。

5. 结论与展望

本文提出了一种基于注意力机制的轻量化模型神经网络，通过设计高效的轻量化注意力模块和自适应特征融合策略，在保持模型轻量化的同时显着提升了图像分类性能。实验结果表明，所提方法在 CIFAR-10 和 ImageNet 数据集上较现有轻量化网络取得了更好的精度–效率平衡。

未来工作分布后续几个方面继续深入研究：神经架构搜索：结合自动化架构搜索技术，进一步优化网络结构设计；多任务学习：探索将所提方法进行分割目标检测、追踪等其他任务；硬件协同优化：针对特定硬件平台进行定制化优化，进一步提升部署效率；知识补充：结合教师–学生网络框架，利用大模型知识训练轻量化模型[5]。

本文提出的轻量化焦点网络为移动端深度学习应用提供了新的解决思路，具有重要的理论意义和应用。

参考文献

[1] 何凯, 张晓, 任胜, 等. 深度残差学习在图像识别中的应用[C]//IEEE 计算机视觉与模式识别会议论文集. 2016: 770-778.

[2] Howard, A.G., Zhu, M., Chen, B., *et al*. (2017) Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications. https://arxiv.org/abs/1704.04861

[3] 胡建军, 沈玲, 孙刚. 挤压激励网络[C]//IEEE 计算机视觉与模式识别会议论文集. 2018: 7132-7141.

[4] Jacob, B., Kligys, S., Chen, B., *et al*. 用于高效整数算术推理的神经网络量化和训练[C]//IEEE 计算机视觉与模式识别会议论文集. 2018: 2704-2713.

[5] Hinton, G., Vinyals, O. and Dean, J. (2015) Distilling the Knowledge in a Neural Network. https://arxiv.org/abs/1503.02531