摘要

近年来，基于深度卷积神经网络与视觉 Transformer 技术的视觉识别模型已取得显著进展。然而，这些模型的应用正面临着参数规模快速增长与计算资源受限之间的矛盾。这一矛盾体现在神经网络架构的算子、模块、结构三个递进层面。首先，在算子层面，自注意力机制的计算复杂度随输入分辨率呈平方级增长，在处理高分辨率图像时，其密集计算对硬件算力提出了严峻挑战。其次，在模块层面，深度神经网络存在显著的参数冗余，且现有的剪枝、参数共享等压缩方案难以兼顾底层硬件的实际加速效果与参数的充分利用。最后，在结构层面，依赖人工经验设计网络结构的效率不高，难以保证在庞大搜索空间中找到最优解，且现有的神经网络搜索方法在处理复杂拓扑结构时泛化能力不足。为突破上述瓶颈，本文面向视觉识别任务，从算子、模块和结构三个层面对视觉神经网络架构进行设计与优化，分别提升了视觉识别的特征融合能力、参数利用效率和自动化结构搜索效率。本文的主要贡献如下：

在算子层面，针对全局注意力复杂度高的问题，本文提出了基于全局特征融合的高效算子设计方法。该方法旨在融合视觉 Transformer 的长程建模优势与卷积神经网络的推理效率优势，构建高性能、低延迟的新型卷积神经网络。该方法利用矩阵分解技术构建超大卷积核，在纯卷积框架下以线性复杂度实现类似自注意力的全局感受野，提升了卷积神经网络对长距离依赖的建模能力；同时，构建了分支门控单元，利用乘法门控机制增强通道间及空间维度上的非线性交互与动态特征筛选能力。实验表明，该方法在 ImageNet 分类任务上精度达到 84.6%，超越了同期的卷积神经网络和视觉 Transformer 网络，并在分割和检测等下游任务中能够有效兼顾精度与推理速度，为高效算子设计提供了更优的解决思路。

在模块层面，针对参数冗余度高的问题，本文提出了基于层剪枝和复用的高效参数压缩方法。该方法从剔除冗余和参数复用两个互补维度展开。在剔除冗余方面，针对神经网络的深度冗余现象，本文提出基于扁平化的 Transformer 层剪枝方法，将相似度高的相邻模块合并，随后利用细粒度通道剪枝精确识别并剔除冗余参数。在参数复用方面，针对多模态大模型的高内存占用问题，本文提出了基于特征对齐的层递归复用方法，通过多模态递归连接器对齐跨模态特征。实验结果表明，扁平化方法在保留 90% 零样本泛化性能的同时，实现了 1.26 倍的响应速度提升和 1.27 倍的吞吐量提升。层递归复用方法在不增加物理参数总量的前提下，通过多轮迭代达到了 62.6% 的平均精度，超越了以往的递归模型方法。

在结构层面，针对人工设计难以保证结构最优的问题，本文提出了基于图表示优化的高效结构搜索方法。该方法重新审视了神经网络有向无环图的拓扑结构，指出以

往研究忽视了兄弟节点在捕捉并行分支特征中的价值。为此，本文设计了邻接-兄弟多头注意力机制，通过注意力掩码显式聚合局部邻居与兄弟节点信息，增强了模型对复杂拓扑结构的建模能力；同时提出了双向图同构前馈网络，利用双向图卷积增强模型对有向图拓扑结构的同构判别能力，摆脱了对复杂位置编码的依赖。在神经网络结构搜索测试基准上的实验表明，该神经预测器在准确率排序与底层硬件延迟预测任务上超越现有主流方法 5-10%，为模型精准评估与排序提供了可靠且高效的自动化代理工具。

为验证上述方法在复杂真实场景中的有效性，本文构建了基于高效神经网络的自动驾驶目标检测系统。自动驾驶场景的目标检测具有运动状态复杂、目标尺度变化大以及响应延迟低等特点，对视觉神经网络的计算效率提出了很高的要求。为此，该系统深度集成了本文提出的高效算子、参数压缩与复用模块以及神经网络结构搜索方法。在车载边缘计算硬件上的部署与测试结果表明，本文提出的方法能够在保持极低内存占用与推理延迟的前提下，有效提升复杂交通场景中关键目标的定位与检测精度。

综上所述，本文面向视觉识别任务，在深度神经网络的算子、模块和结构三个层面开展了系统研究，成功缓解了参数规模快速增长与计算资源受限之间的矛盾。本文的研究不仅在科学层面推进了轻量化深度神经网络理论的发展，也为深度神经网络在端侧设备上的实时部署和应用提供了技术支撑。

关键词：视觉识别，神经网络架构，模型剪枝，参数共享，神经网络结构搜索

Research on Efficient Deep Neural Network Architectures for Visual Recognition

Ruihan Xu (Computer Application Technology)
Supervised by Tenured A.P. Shiliang Zhang

ABSTRACT

In recent years, visual recognition models based on deep Convolutional Neural Networks (CNNs) and Vision Transformers have achieved remarkable progress. However, their practical application is increasingly constrained by the tension between the rapid growth of model parameters and limited computational resources. This tension is manifested in three progressive levels of neural network architecture, namely operators, modules, and structures. First, at the operator level, the computational complexity of the self-attention mechanism grows quadratically with the input resolution, posing severe challenges to hardware when processing high-resolution images. Second, at the module level, deep neural networks contain considerable parameter redundancy, and existing compression schemes such as pruning and parameter sharing struggle to simultaneously achieve practical hardware acceleration and sufficient parameter utilization. Third, at the structure level, network architectures that heavily rely on manual design are inefficient and make it difficult to guarantee optimal solutions in a vast search space, while existing neural architecture search methods show limited generalization when handling complex topologies. To address these bottlenecks, this thesis focuses on visual recognition tasks and investigates efficient neural network architectures from the three progressive levels of operators, modules, and structures, thereby improving feature fusion capability, parameter utilization efficiency, and automated structure search efficiency. The main contributions of this thesis are summarized as follows:

At the operator level, to address the high complexity of global attention, this thesis proposes an efficient operator design method based on global feature fusion. The proposed method aims to combine the long-range modeling advantages of Vision Transformers with the inference efficiency of CNNs, and constructs a novel CNN architecture with high performance and low latency. By using matrix decomposition, ultra-large convolution kernels are built to realize a self-attention-like global receptive field with linear complexity under a pure convolutional framework, thereby improving the ability of CNNs to model long-range dependencies. In ad-

dition, a branch gating unit is introduced to strengthen nonlinear interactions across channels and spatial dimensions, as well as dynamic feature selection, through a multiplicative gating mechanism. Experimental results show that the proposed method achieves 84.6% accuracy on ImageNet classification, outperforming contemporary CNNs and Vision Transformers. It also effectively balances accuracy and inference speed in downstream tasks such as segmentation and detection, providing a more effective solution for efficient operator design.

At the module level, to address the problem of high parameter redundancy, this thesis proposes an efficient parameter compression method based on layer pruning and reuse. This method is developed from two complementary perspectives: redundancy removal and parameter reuse. For redundancy removal, in view of the depth redundancy in neural networks, a flattening-based Transformer layer pruning method is proposed, which merges adjacent modules with high similarity and then uses fine-grained channel pruning to accurately identify and remove redundant parameters. For parameter reuse, to address the high memory footprint of multimodal large models, a layer-recursive reuse method based on feature alignment is proposed, in which a multimodal recursive connector is designed to align cross-modal features. Experimental results show that the flattening method improves response speed and throughput by factors of 1.26 and 1.27, respectively, while retaining 90% of zero-shot generalization performance. Without increasing the total number of physical parameters, the layer-recursive reuse method achieves an average accuracy of 62.6% through multiple rounds of iteration, outperforming previous recursive model methods.

At the structure level, to address the problem that manual design cannot guarantee optimal structures, this thesis proposes an efficient structure search method based on graph representation optimization. This method revisits the topology of neural network directed acyclic graphs and points out that previous studies overlooked the value of sibling nodes in capturing parallel branch features. To this end, an adjacency-sibling multi-head attention mechanism is designed to explicitly aggregate information from local neighbors and sibling nodes through attention masks, thereby enhancing the model's capability to represent complex topological structures. Meanwhile, a bidirectional graph isomorphism feed-forward network is proposed to improve the model's ability to distinguish isomorphic directed graph topologies using bidirectional graph convolutions, thus avoiding reliance on complex positional encodings. Experiments on neural architecture search benchmarks show that the proposed neural predictor outperforms existing mainstream methods by 5-10% in both accuracy ranking and underlying hardware latency prediction tasks, providing a reliable and efficient automated proxy for accurate evaluation and ranking of models in heterogeneous search spaces.

To verify the effectiveness of the proposed methods in complex real-world scenarios, this thesis further constructs an autonomous driving object detection system based on efficient neural networks. Object detection in autonomous driving is characterized by highly dynamic changes, a wide range of object scales, and millisecond-level latency requirements, which impose stringent demands on the computational efficiency of visual neural networks. To this end, the system deeply integrates the efficient operators, parameter compression and reuse modules, and neural architecture search methods proposed in this thesis. Deployment and testing results on in-vehicle edge computing hardware demonstrate that the proposed methods can effectively improve the localization and detection accuracy of key objects in complex traffic scenarios while maintaining extremely low memory consumption and inference latency.

In summary, this thesis conducts systematic research on visual recognition tasks from the operator, module, and structure levels of deep neural networks, and successfully alleviates the tension between the rapid growth of model parameters and limited computational resources. This work not only advances the development of lightweight deep neural network theory at the scientific level, but also provides technical support for the real-time deployment and application of deep neural networks on edge devices.

KEY WORDS: Visual Recognition, Neural Network Architecture, Model Pruning, Parameter Sharing, Neural Architecture Search