http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2025.0313

面向不规则结构的低空感知模型剪枝算法

杨子，庄连生*
(中国科学技术大学 网络空间安全学院，合肥 230026)

摘 要：深度学习在低空智能感知中性能优越，但终端计算资源有限，难以部署大型神经网络。剪枝技术通过删除冗余参数降低复杂度，提升实时性，但现有算法多针对规则结构模型，难以处理神经架构搜索(NAS)得到的不规则结构。提出一种基于依赖图的不规则结构剪枝方案，以优化模型效率。通过计算图的结构解析技术，自动识别低空感知模型中特有的多层级连接模式；构建高效的参数分组模块，精确锁定可完整移除的最小功能单元；采用跨层对比的全局优化策略，实现对同构子结构的统一评估与剪枝。相较于现有算法，所提算法可对不规则网络架构实施全局精准压缩，特别适用于需要处理复杂空间关系的低空感知场景。实验结果表明：所提不规则结构剪枝算法在NAS-Bench101模型及ResNet系列网络均能有效识别出多分支连接、残差连接、拼接等复杂结构，并完成分组剪枝，在CIFAR-10数据集以50%的压缩比例对模型轻量化，识别准确率下降不超过1%。

关 键 词：低空感知；卷积网络；轻量化模型；模型剪枝；不规则网络架构
中图分类号：V221+.3; TB553
文献标志码：A **文章编号**：1001-5965(2026)09-3136-10

低空智能感知是基于人工智能、多源传感融合和协同计算等技术，在低空空域实现对动态目标、环境状态及飞行器行为的全天候、高精度、全方位实时监测与智能分析的综合技术体系。各种新型低空运载工具的广泛应用，对低空环境感知和低空安全提出了更高要求。深度神经网络在低空动态障碍物的实时感知[1-4]、轨迹预测[5-7]等关键场景中展现出卓越性能。由于计算资源受限，传统大规模神经网络往往难以直接部署到低空飞行器上。模型剪枝通过移除神经网络中的冗余结构，将部分参数置为非激活状态，达到简化模型结构、降低计算资源消耗的目的，使神经网络模型在飞行器终端的部署成为可能，对于提高系统实时环境感知能力至关重要，有助于推动低空智能感知技术的产业化进程[8-10]。

依据剪枝粒度的差异，模型剪枝可分为非结构化剪枝和结构化剪枝2类[11]。非结构化剪枝主要通过剪除滤波器层内或跨层的连接，从而有效降低网络存储成本。然而，这种方法在降低深度神经网络计算成本方面效果有限。1989年，Lecun等提出了最佳脑损伤[12]理论，运用二阶导数信息平衡网络复杂度和训练误差，进而通过移除网络中不重要的权重实现模型优化。Han等开发了基于L_1范数的权重剪枝算法[13]，通过评估神经元连接权值的范数大小，删除低于预设阈值的连接。Yang等[14]创新性地将层级能量消耗作为剪枝顺序的判定标准，通过加权稀疏投影和输入遮蔽技术量化能耗指标，并将能耗预算整合为网络训练的约束条件。Gong等[15]提出了一种快速可控的训练后稀疏化框架，通过引入可微分桥函数和可控优化目标，解决了传统训练后剪枝算法因忽略分层稀疏率分配导致的精度显著下降问题。需要指出的是，尽管非结构化剪枝能够有效压缩模型规模，但其产生的非结构化稀疏特性与现有基础线性代数子程序库的兼容性较差，导致其难以充分发挥硬件加速优势。

收稿日期：2025-05-19；录用日期：2025-05-30；网络出版时间：2025-06-05 14:26
网络出版地址：link.cnki.net/urlid/11.2625.V.20250605.1115.003
*通信作者. E-mail：yangzi@mail.ustc.edu.cn

引用格式：杨子，庄连生. 面向不规则结构的低空感知模型剪枝算法[J]. 北京航空航天大学学报，2026，52(9)：3136-3145.
Yang Z，Zhuang L S. Low-altitude perception models pruning algorithm with irregular structures[J]. Journal of Beijing University of Aeronautics and Astronautics，2026，52(9)：3136-3145 (in Chinese).

相比之下, 结构化剪枝是一种硬件友好型剪枝方式, 该技术以深度神经网络中滤波器所在的通道为剪枝对象, 通过评估通道重要性实现结构化修剪。因此, 结构化剪枝在降低存储开销的同时, 也能有效提升网络推理速度。相较于非结构化剪枝, 结构化剪枝能够生成结构化的稀疏模型, 既减少设备存储成本, 又可实现硬件层面的实际加速效果。结构化剪枝的核心挑战在于如何有效筛选重要滤波器。现有研究提出了多种滤波器重要性评估算法[16-19]: Li 等[16] 采用卷积核权重绝对值计算方法, 通过计算滤波器的 L_1 范数筛选目标; Hu 等[18] 提出的剪枝算法从另一个角度进行重要性评估, 通过分析 ReLU 激活函数的输出特性评估神经元重要性; Liu 等[19] 开发了基于批归一化参数的剪枝策略, 利用批归一化层中缩放因子的 L_1 范数作为通道重要性指标; Lin 等提出跨层排序-k-倒数最近滤波器 (cross-layer ranking-k-reciprocal nearest filter, CLR-RNF) 方法[20], 设计了基于推荐机制的评估体系, 运用 k 近邻算法为每个滤波器构建相似候选集, 再通过多组候选集的交集运算识别关键滤波器。

随着低空环境应用需求的升级, 剪枝技术正面临严峻挑战。为适应复杂多变的低空感知任务, 现代网络架构已突破传统层级堆叠模式, 演化出动态连接和多分支拓扑等新型结构[21], 这些创新显著增加了模型轻量化的实施难度。在低空智能感知场景中, 模块间高度耦合的特征传递机制使任何参数调整都可能破坏空间关系推理链条, 导致目标识别精度的大幅波动。这些非常规连接结构往往由算法演化生成, 而非人工设计, 导致其可能具有任意拓扑结构和近乎无限的连接复杂度。面对此类网络的剪枝需求时, 常规的逐层裁剪策略已不再适用, 因为随意切断连接会引起模块输入输出维度的突变, 破坏相邻层间的结构兼容性, 最终导致模型失效。现有剪枝算法通常针对固定网络架构设计, 并在标准基准网络上进行验证[22-24]。当应用于不规则网络架构时, 这些方法往往需要依赖繁琐的手动调整(如通道对齐或跳跃连接优化), 或引入额外补偿机制(如动态重参数化技术), 这种适应性缺陷严重制约了模型轻量化算法的实际应用效率。现有工作针对不同类型的神经网络存在着不同的轻量化算法。例如, 针对卷积网络, 研究人员提出了深度可分离卷积方法[25]; 针对循环神经网络, 研究人员提出了时序动态剪枝算法[26]; 针对 Transformer 网络, 研究人员提出了 token 稀疏化机制[27]。但这些方法的设计与对应的网络架构类型密切相关, 缺乏通用性。本质上, 这些方法都是针对规则网络架构而设计的, 无法解决不规则网络架构的轻量化问题。

针对以上挑战, 本文提出基于依赖图的不规则结构剪枝算法, 从依赖关系分组的角度对结构化剪枝算法进行重构, 以实现结构化剪枝算法在不规则结构网络上的适配。通过依赖分析模块对网络计算图进行解析, 动态追踪计算图的数据流路径, 建立跨层连接的依赖子图。第 2 个参数分组模块通过动态追踪前向计算图中的数据流与控制流依赖关系, 自动识别可被完整移除的最小功能单元, 避免剪枝引发的逻辑断裂问题。为进一步提升全局压缩效率, 增加了同构评估剪枝模块, 通过设计跨网络层的子图同构检测算法, 实现对相似拓扑单元的重要性评估与批量移除。与已有算法相比, 本文算法从依赖图分析的角度突破传统剪枝的层内限制, 并通过拓扑单元的动态识别与联合优化, 有效解决了不规则网络中的参数依赖问题。

1　面向不规则结构的模型剪枝算法

针对现有结构剪枝算法依赖人工设计分组方案、难以处理复杂依赖关系的局限性, 本文提出基于依赖图建模的通用结构剪枝框架, 通过构建神经网络层间的拓扑依赖模型实现任意架构的自动压缩, 如图 1 所示。该方案的总体思路在于引入依赖图显式表征跨层计算关系, 具体包含 3 个核心递进模块: ①依赖分析模块, 通过追踪数据流路径, 构建依赖关系模型; ②参数分组模块, 基于图连通性分析自动识别可独立移除的最小子结构; ③评估剪枝模块, 提出组级联合评估准则, 实现子图结构的全局批量剪枝。

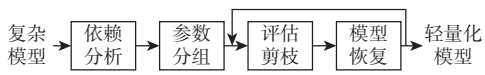


图 1　面向不规则结构的模型剪枝算法框架

Fig. 1　Frame of models pruning algorithm with irregular structures

1.1　依赖分析模块

依赖分析模块的核心任务是识别神经网络中各层之间的依赖关系, 以构建结构化的依赖图模型, 为后续自动化剪枝提供拓扑关系基础。本节解释依赖图模型的构建原理, 即图的节点代表网络中的部分。一种直观的初步思路是: 以一整层的参数作为一个基本单元, 构建不同层之间的依赖关系图模型。然而, 在实际剪枝过程中, 若采用层级别的构建方式, 所构建出的依赖关系图会产生冲突, 因为一些基本网络模块(如全连接层)可能有 2 种不同的剪枝模式, 如剪枝二维权重矩阵的 $\boldsymbol{w}[k,:]$ 和 $\boldsymbol{w}[:,k]$, 这 2 种方案分别会压缩输入和输出的维

度。此外，网络还包含非参数化操作，如残差连接，这也会影响层之间的依赖性。为解决上述问题，本文使用一种新的表示法，将网络$F(x;\boldsymbol{w})$分解为更精细、更基本的分量，表示为$F=\{f_1,f_2,\cdots,f_L\}$，其中，每个分量f_i表示参数化层（如卷积）或非参数化运算（如残差加法）。本文不在层级别建模关系，而是专注于层的输入和输出之间的依赖关系，即将分量f_i的输入和输出分别表示为f_i^-和f_i^+。对于任何神经网络层或模块，最终分解都可以形式化为$F=\{f_1^-,f_1^+,f_2^-,f_2^+,\cdots,f_L^-,f_L^+\}$。这种表示方法有助于简化依赖关系建模，并对同一层使用不同的剪枝方案。利用该符号可以将神经网络重新用公式表示，其中，可以区分2种主要类型的依赖关系，即层间依赖关系和层内依赖关系，表示为

$$(f_1^-,f_1^+)\leftrightarrow(f_2^-,f_2^+)\cdots\leftrightarrow(f_L^-,f_L^+) \tag{1}$$

式中："↔"表示2个相邻层之间的连接性，即一个层的输出作为另一个层的输入。

依赖关系成立的条件总结为2种：①层间依赖关系，仅当i层和j层前后相连时，存在依赖关系$f_i^-\Leftrightarrow f_j^+$，这一条件表示为$f_i^-\leftrightarrow f_j^+$；②层内依赖关系，仅当$f_i^-$和$f_i^+$拥有相同的剪枝模式时，存在依赖关系$f_i^-\Leftrightarrow f_i^+$，这一条件表示为$\mathrm{sch}(f_i^-)=\mathrm{sch}(f_i^+)$。

层间的参数依赖显然由网络前向推理的路径依赖所决定。对于相连接的层$f_i^-\leftrightarrow f_j^+$，两者存在依赖关系，因为$f_i^-$和$f_j^+$分别代表先前层的输出和后继层的输入，两者对应于网络的同一组中间特征。同一层的参数之间也可能存在依赖性，即层内的参数依赖，如果某一层的输出和输入具有层内依赖，那么在移除参数时必须同时剪枝单个层的输入和输出。

层内依赖模式如图2所示。这种依赖性是由该层的计算方式决定的，许多网络层满足这一条件。例如，批量归一化层输入和输出被剪枝时，需要移除的参数相同，表示为$\mathrm{sch}(f_i^-)=\mathrm{sch}(f_i^+)$。相反，卷积层对其输入和输出具有不同的剪枝方案，即$\boldsymbol{w}[:,k,:,:]\neq\boldsymbol{w}[k,:,:,:]$，导致$\mathrm{sch}(f_i^-)\neq\mathrm{sch}(f_i^+)$，如图2所示，在这种情况下，卷积层的输入和输出之间不存在依赖性。

给定上述规则，本文可以形式化地建立依赖关系：

$$D(f_i^-,f_j^+)=\mathbb{1}[f_i^-\leftrightarrow f_j^+]\vee \mathbb{1}[i=j\wedge\mathrm{sch}(f_i^-)=\mathrm{sch}(f_j^+)] \tag{2}$$

式中：$\mathbb{1}$为指示函数；$\wedge$为逻辑与运算。

式(2)中，右侧第1项负责捕捉网络连接产生的跨层关联，第2项则用于分析层输入与输出之间因剪枝模式相同而具有的层内依赖性。依赖矩阵

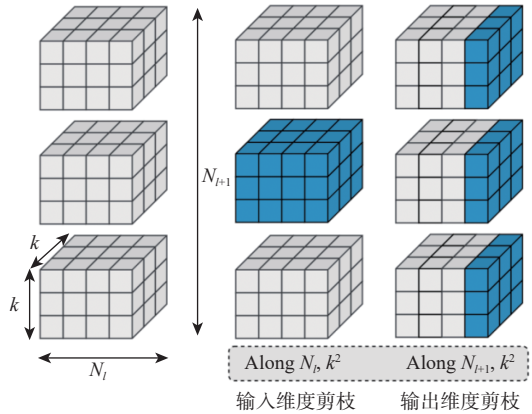


图2 层内依赖模式示意

Fig. 2 Illustration of inter-layer dependency patterns

具有对称特性，严格遵循$D(f_i^-,f_j^+)=D(f_j^+,f_i^-)$的对称关系。基于这种对称特性，本文可以通过系统遍历所有输入输出配对，来完整精确地构建完整的依赖拓扑模型。图3直观展示了带有残差连接的卷积网络块所对应的依赖关系图构建过程。

1.2 参数分组模块

参数分组模块的任务是将具有依赖关系的网络子结构进行分组，确保这些子结构在剪枝时能够被同时处理。为此，本文基于依赖图进行分组，通过将参数划分问题转化为依赖图的连通分量划分问题，将相互依赖的子结构划分为同一组。具体而言，依赖图在构建时，每一个节点都与其对应的参数进行关联，当依赖图的2个节点连通时，其对应的参数也被划分到同一个参数分组。对于不同的依赖类型，寻找所需要分组的参数具有不同的处理方法。将常见的依赖关系拆解为若干个基本类型，如图4所示，以展开对参数分组模块的详细描述。

从一个无分支神经网络开始解释参数分组方法，如图4(a)所示，该网络由3个连续层组成，网络参数由二维权重矩阵$\boldsymbol{w}_l$、$\boldsymbol{w}_{l+1}$和$\boldsymbol{w}_{l+2}$组成。这种简单的神经网络可以通过去除神经元进行结构化剪枝。在这种情况下，很容易发现参数之间出现了一些依赖关系，表示为$\boldsymbol{w}_l\Leftrightarrow\boldsymbol{w}_{l+1}$，表示$\boldsymbol{w}_l$和$\boldsymbol{w}_{l+1}$必须同时被剪枝。具体而言，为剪枝桥接$\boldsymbol{w}_l$和$\boldsymbol{w}_{l+1}$的第$k$个神经元，$\boldsymbol{w}_l[k,:]$和$\boldsymbol{w}_{l+1}[k,:]$都将被移除。无分支网络依赖模式如图5所示，其中，蓝色部分表示被剪枝的特征图及其对应的参数位置。

残差依赖如图6所示。对带有残差连接的残差块进行剪枝时，残差块第2层中待剪枝的滤波器必须由残差的剪枝结果决定，而仅有残差块的第1层可以无限制地进行剪枝。类似的简单的依赖关系可能会相互嵌套或组合，从而在复杂拓扑结构的网络中形成更为复杂的模式。

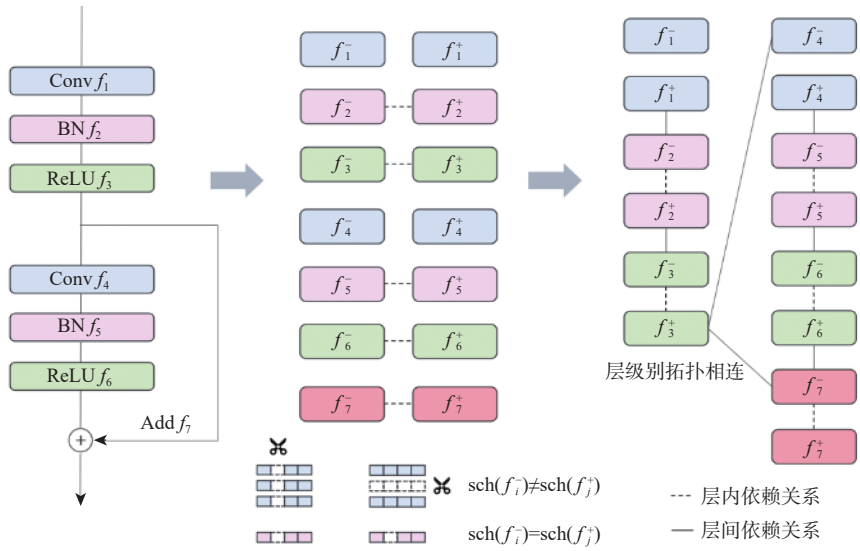


图 3　依赖关系图构建

Fig. 3　Construction of a dependency relationship graph

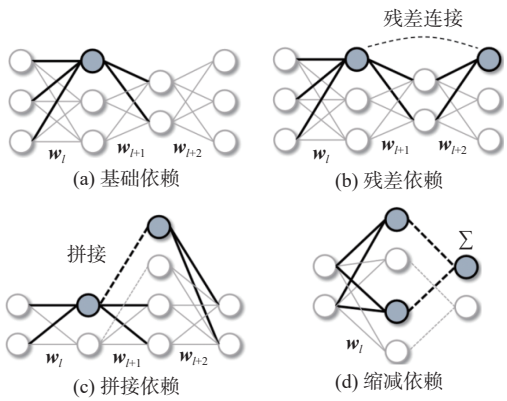


图 4　不同结构中具有相互依赖性的分组参数

Fig. 4　Grouped parameters with dependency in different structures

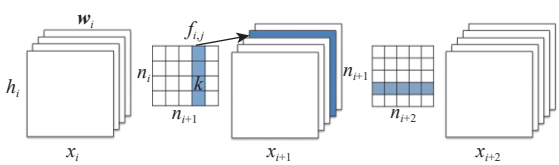


图 5　无分支网络依赖模式示意

Fig. 5　Dependency patterns diagram in branchless networks

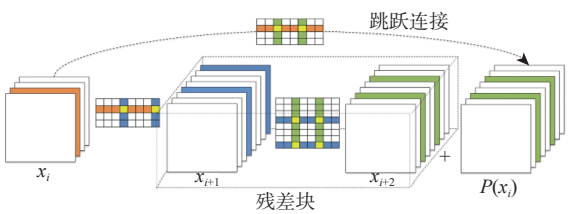


图 6　残差网络依赖模式示意

Fig. 6　Dependency patterns diagram in residual networks

除残差连接导致的额外依赖关系外, 不规则结构的神经网络中还存在一些其他依赖关系需要处理。拼接依赖在神经网络架构中较为常见, 尤其是在一些需要融合不同特征信息的模型模块中, 如图 7 所示。以 U-Net 网络为例, 在其编码器-解码器结构中, 当从编码器向解码器传递特征时, 会存在拼接操作。这种拼接依赖具体表现为: 不同层的特征图在通道维度上直接相连, 在拼接操作后紧接层的输入维度将由拼接操作前若干层的输出维度共同决定。从数学角度看, 若有 2 个特征图 $\boldsymbol{A}$ 和 $\boldsymbol{B}$, 其维度分别为 (N,C_1,H,W) 和 (N,C_2,H,W), 经过拼接操作后, 得到的新特征图 $\boldsymbol{C}$ 的维度为 (N,C_1+C_2,H,W), 新特征图 $\boldsymbol{C}$ 中的每个通道分别来源于 $\boldsymbol{A}$ 和 $\boldsymbol{B}$, 这种依赖关系使模型能够综合不同层级的特征, 提升对复杂目标的分割或识别能力。

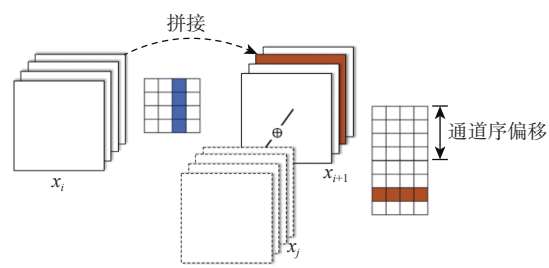


图 7　拼接依赖模式示意

Fig. 7　Dependency pattern diagram with concatenation

缩减依赖也是现代神经网络中不容忽视的一种依赖形式。在如 SqueezeNet[28] 这类模型中, 为降低计算量和模型参数, 会采用挤压模块, 该模块由挤压层和扩展层组成, 挤压层通过使用 1×1 的卷积核来缩减通道数, 从而形成一种缩减依赖。具体而言, 假设输入特征图的通道数为 C_{in}, 挤压层的输出通道数为 C_{squeeze} ($C_{\text{squeeze}} \ll C_{\text{in}}$), 挤压层通过对输入特征图在空间维度上进行卷积操作, 将每个空间位置的 C_{in} 个通道特征进行线性组合, 得到 C_{squeeze} 个通道的输出。这种依赖使后续扩展层的计算量显著

降低，因为扩展层是基于挤压层缩减后的通道数进行操作的。在该过程中，挤压层输出的特征图对后续扩展层的输入有直接的依赖关系。

对于上述各类依赖模式，参数分组模块将关联其对应的参数，并在计算依赖图的最大连通子分量后，将对应的参数划分为一组，完成所有具有依赖关系的参数分组，以便后续的移除。

1.3 评估剪枝模块

本节评估剪枝模块衡量参数的重要性，并实现冗余参数的移除。与现有算法有所不同，该算法以耦合参数组作为衡量参数重要性的基本单元，这就需要应对以下2个关键问题：①单个耦合参数组中往往包含来自不同层的参数，而不同类型的网络层在参数特征和计算拓扑结构上存在显著的多样性，使传统方法中衡量参数重要性的计算方式难以直接应用；②网络内不同层通常呈现出不同程度的冗余性，因此，必须确保不同参数组之间具备可比性，以便进行有效的重要性评估。为解决上述问题，本文算法提出的主要策略是依据子结构的拓扑逻辑及参数配置情况，将子结构聚类到不同的同构群中。在这种聚类方式下，同构子结构的参数之间呈现出一一对应的关系，且不同参数参与计算的方式完全一致，如图8所示。同一同构群内的所有子结构均具有相同的架构设计、参数规模及计算图。

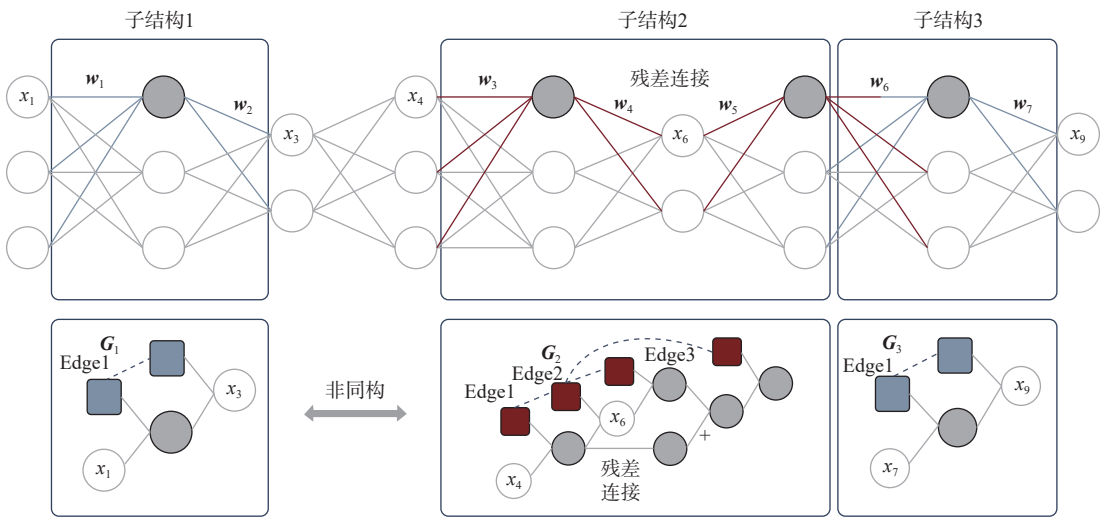


图8 神经网络中的同构子结构

Fig. 8 Isomorphic substructures in a neural network

对于通过前述方法得到的依赖图$\boldsymbol{D}_1$，将每个节点由三元组$(\boldsymbol{W}_1,d,k_1)$进行标记，包含参数矩阵$\boldsymbol{W}_1$、轴d及索引k_1。标签分配遵循以下原则：当且仅当参数矩阵$\boldsymbol{W}'$源自相同类型的网络层（如线性层或卷积层）且具有相同的剪枝维度$d=d'$时，节点$(\boldsymbol{W}_1,d,k_1)$和$(\boldsymbol{W}',d',k')$才被赋予相同的标签。允许索引k不同，因为在相同参数矩阵中，沿第d维切分的权重向量具有一致的齐次性，可视为同构单元。按照上述方法确定顶点标签后，同构参数组要求$\boldsymbol{G}_1$与$\boldsymbol{G}_2$的连接具备同构性。如图8所示，图中的3个子结构分别对应图$\boldsymbol{G}_1$、$\boldsymbol{G}_2$和$\boldsymbol{G}_3$，其中，$\boldsymbol{G}_1$和$\boldsymbol{G}_2$各含单一边，$\boldsymbol{G}_3$则包含3条边。若按计算图的执行顺序检查边连接，可自然观察到$\boldsymbol{G}_1$与$\boldsymbol{G}_3$的同构性，因其共享相同的拓扑结构。为形式化描述该过程，将图$\boldsymbol{G}$中的边按前向传播执行顺序排序，即可通过式(3)判定同构性：

$$I(\boldsymbol{G},\boldsymbol{G}'))=\mathbb{1}\{|\boldsymbol{G}|=|\boldsymbol{G}'|\wedge\text{label}(V_i)=\text{label}\left(V_i'\right),i\in\{1,2,\cdots,|\boldsymbol{G}'|\}\} \tag{3}$$

式中：$|\boldsymbol{G}|$、$|\boldsymbol{G}'|$为图的边数。

首项条件确保图的规模一致性，次项要求所有顶点标签严格匹配。在参数组剪枝阶段，首先，通过定义重要性函数量化各参数组的显著性。具体而言，算法会对一组具有相同图拓扑的子结构进行检查。对于每个图的重要性进行独立估计，进而得到一个多维重要性向量，这一向量的生成有助于直接对重要性进行排序，根据排序得分删除不重要的子结构。对于其他同构群，可依据相同的原理实施剪枝操作，从而实现对网络模型的有效压缩和优化。例如，采用幅度剪枝时，重要性函数为参数组的L_2范数；采用泰勒剪枝时，则结合梯度与参数的乘积进行二阶近似。其次，需排除对网络功能完整性敏感的层，如分类网络的最后一层全连接层，以防止输出维度缩减导致任务失效。初始化剪枝器时需指定目标稀疏度、迭代次数及示例输入（用于依赖关系分析）。最后，在每次剪枝迭代中，剪枝器通过动态调整通道数实现参数组的移除，并通过微调恢复模型性能。

1.4　模型恢复

模型经过剪枝后, 其性能通常会有所下降, 这主要是因为移除部分权重破坏了原有的知识表示和模型结构。因此, 必须对剪枝后的模型进行微调, 使模型在剪枝后的新结构基础上, 重新适应并优化其参数配置。本文使用原数据集对剪枝后的模型进行微调, 使用原模型 1/10 的学习率进行微调后, 回到评估剪枝模块进行迭代剪枝。

2　实验结果与分析

为验证算法的有效性, 借助神经架构搜索(neural architecture search, NAS)方法在不同数据集上搜索最佳不规则网络架构, 对其进行模型压缩, 并比较压缩前后模型的性能。

2.1　实验设置

本节在 CIFAR-10、CIFAR-100 及 ImageNet 基准数据集上通过实验证明了本文压缩算法的有效性。通过将所有训练图像(不包含测试图像)作为训练数据, 对模型进行训练或微调后, 对测试图像分类, 得到最终的分类准确率。在数据处理方面, 按照前人研究的惯例, 使用通道平均值和标准差对输入数据进行归一化, 并使用包含移位、镜像的标准数据增强方案。本文算法在 Pytorch 平台上实现, 并对当前具有一流性能的卷积网络 ResNet 进行剪枝。为获得每个网络的基准精度, 从头开始训练 ResNet 模型, 并遵循与标准 ResNet 训练过程相同的预处理和超参数。本文对 NAS 生成的模型系列 NAS-Bench101 进行压缩测试, 采用的不规则网络架构来源于 NAS-Bench101 网络架构系列。NAS-Bench101 是专门用于神经网络结构搜索的基准数据库, 其核心设计是通过组合标准模块来构建完整网络。该系列将网络拆分为多个重复的单元模块(称为细胞), 每个细胞由 3～7 个计算节点通过不同路径连接构成。节点间的连接线可选择 3×3 卷积、1×1 卷积或 3×3 最大池化等基础计算操作。通过调整节点数量、连接方式和操作类型, 该基准库共预定义了 42.3 万种有效网络结构, 其深度跨度从 9 层到 277 层, 模型参数规模从 20 万到 3 730 万不等。不同结构的计算量差异显著, 处理单张 CIFAR-10 图像所需浮点运算量最低仅 3×10^7 次, 最高可达 8.9×10^9 次, 展现出丰富的结构多样性。在训练及微调时, 本文均使用以下设置: 使用 Tesla-V100-DGXS-16 GB GPU, 在标准数据集, 同一批处理的图像数量为 128; 使用 Adam 优化器进行优化, 学习率为 0.001, 剪枝加微调共进行不超过 40 轮迭代(一般待压缩网络模型的训练需要数十倍), 并保存具有最佳验证准确率的压缩后模型作为算法的输出。

与绝大部分研究方法一致, 本文利用浮点运算量和参数量衡量网络复杂度。浮点运算量表示网络在推理过程中所执行的浮点运算的总次数, 用于评估网络的计算量; 参数量则表示网络中权重参数的总量, 用于评估网络的规模。本文利用剪枝比率(pruning rate, PR)表示网络在剪枝后的压缩比例, 使用 Top-1 精度指标, 即最高预测概率的类别是否为正确类别, 以评估原始网络和剪枝后网络的性能。

2.2　参数分组可视化

为系统验证神经网络中各层级间的参数依赖关系, 本文将网络依赖矩阵和根据其推理得到的参数分组矩阵进行可视化, 结果如图 9 所示。分组矩阵依据本文所表述的图算法中依赖图推导得出, 其中, $G^*[i, j]=1$ 表示第 i 层与第 j 层属于同一组。从可视化结果可以看出, 无论是 ResNet 还是 NAS-Bench 中, 同一残差块或细胞结构内的各层之间呈现出很强的相关性, 使进行结构化剪枝时形成较大的分组。特别是网络搜索得到的细胞结构内部存在多路径并行的参数依赖, 导致大多数分组跨越超过 4 层, 这与传统链式网络中常见的 2～3 层分组形成鲜明对比。这种现象揭示了 NAS 网络特有的参数耦合特性: 为追求更高的表征能力, NAS 过程会自发形成复杂的跨层连接, 但这种结构特性也使手动剪枝极易破坏网络的完整性。

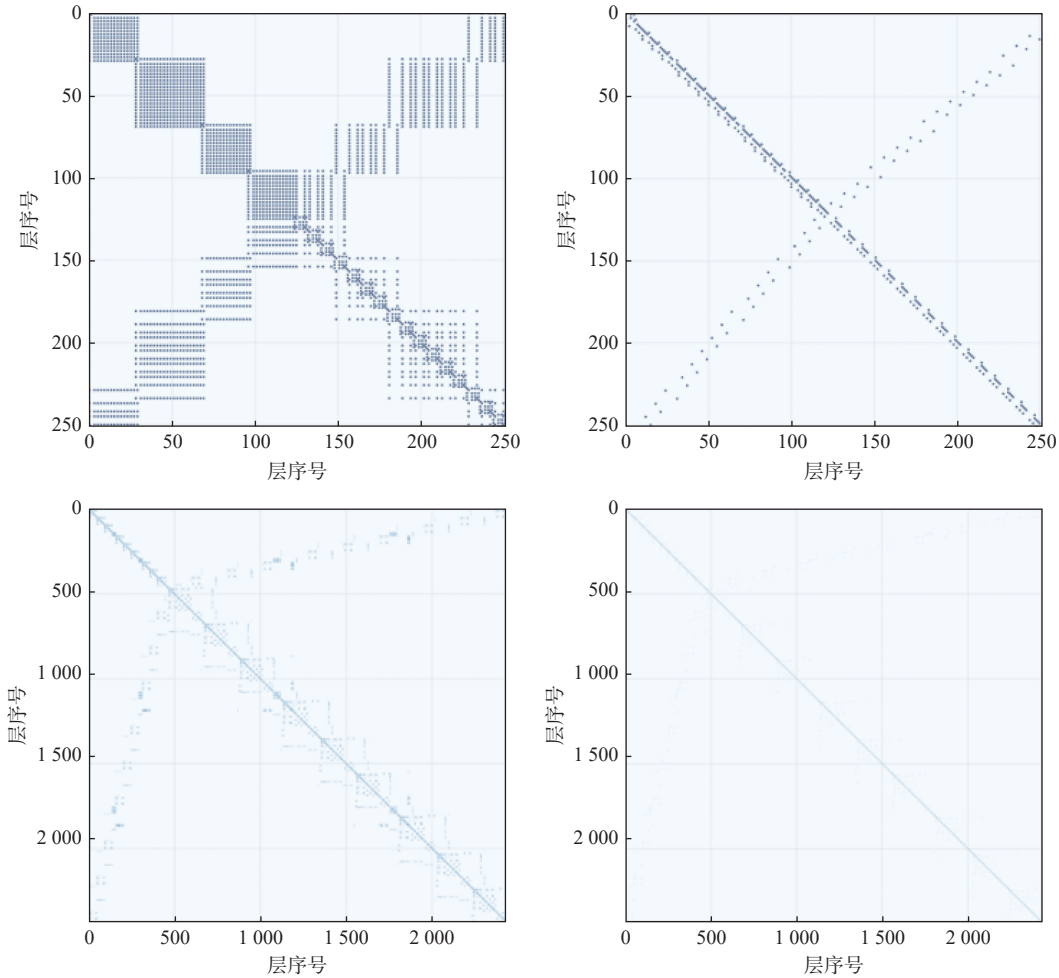
2.3　对比实验

表 1～表 3 中列出了本文算法和其他算法的对比结果。针对 ResNet 网络进行实验, 表 1 展示了使用 50% 的压缩比例在 CIFAR-10 数据集识别任务上压缩 ResNet56 网络的结果, 表 2 展示了使用 50% 的压缩比例在 CIFAR-100 数据集识别任务上压缩 ResNet101 网络的结果。由于模型参数需要成块地被压缩, 因此, 当达到压缩比例目标时, 实际压缩比例可能略大于 50%。本文算法超过了目前的主流模型压缩算法, 在相同的压缩比例设置下取得了最好的识别率。表 3 列出了本文使用 NAS 产生的高性能模型进行压缩的结果。

表 4 展示了不同压缩比例对模型性能的影响。实验基于 CIFAR-10 数据集开展, 通过本文压缩算法生成不同压缩程度的模型, 除目标压缩比例外, 其余参数设置均与基准实验保持一致。在 10%～80% 的压缩比例范围内, 以 10% 为间隔梯度式调节压缩比例, 最终获得 8 组对比实验数据。

2.4　消融实验

通过系统设计的消融实验验证关键算法组



(a) 参数分组矩阵可视化结果 (b) 网络依赖矩阵可视化结果

图 9 分组矩阵可视化

Fig. 9 Group matrix visualization

表 1 在 CIFAR-10 数据集上的识别准确率对比 (ResNet56)

Table 1 Comparison of recognition accuracy rate on CIFAR-10 dataset (ResNet56)

算法	参数量	压缩比例/%	识别准确率/%
不压缩	2.352×10^7	0	94.90
L_1-norm	1.151×10^7	51.06	93.37
L_2-norm	1.151×10^7	51.06	93.86
FPGM	1.151×10^7	51.06	93.64
Taylor	1.151×10^7	51.06	93.80
本文	1.151×10^7	51.06	94.93

表 2 在 CIFAR-100 数据集上的识别准确率对比 (ResNet101)

Table 2 Comparison of recognition accuracy rate on CIFAR-100 dataset (ResNet101)

算法	参数量	压缩比例/%	识别准确率/%
不压缩	4.270×10^7	0	78.30
L_1-norm	2.094×10^7	50.96	74.75
L_2-norm	2.094×10^7	50.96	75.59
FPGM	2.094×10^7	50.96	74.45
Taylor	2.094×10^7	50.96	75.42
本文	2.094×10^7	50.96	76.68

表 3 在 CIFAR-10 数据集上的识别准确率对比 (NAS)

Table 3 Comparison of recognition accuracy rate on CIFAR-10 dataset (NAS)

算法	参数量	压缩比例/%	识别准确率/%
不压缩	4.30×10^6	0	96.97
L_1-norm	2.14×10^6	50.24	95.81
L_2-norm	2.14×10^6	50.24	95.50
FPGM	2.14×10^6	50.24	95.61
Taylor	2.14×10^6	50.24	95.77
本文	2.14×10^6	50.24	95.90

表 4 不同压缩比例下的性能对比

Table 4 Performance comparison at different compression ratios

压缩比例/%	参数量	浮点运算量/次	识别准确率/%
10	2.108×10^7	5.717×10^{10}	95.10
20	1.856×10^7	5.021×10^{10}	95.14
30	1.635×10^7	4.433×10^{10}	94.92
40	1.394×10^7	3.786×10^{10}	95.01
50	1.151×10^7	3.119×10^{10}	94.93
60	9.210×10^7	2.511×10^{10}	94.61
70	7.460×10^6	2.036×10^{10}	94.57
80	4.550×10^6	1.241×10^{10}	94.03

件。在分组策略验证中, 依次测试了无分组独立剪枝、卷积层分组剪枝和全参数分组剪枝 3 种模式, 保持其他实验条件一致; 在稀疏性调控实验中, 采用相同的总体压缩比例约束, 对比了固定比例均匀剪枝和动态分配全局剪枝 2 种策略。所有实验均在标准数据集和基准模型上进行, 以确保实验结果的可靠性。

网络各部分的稀疏性设置是剪枝过程中的重要因素, 其决定了剪枝后神经网络的最终结构。为探究稀疏性设置对结构化剪枝的影响, 以 NAS-Bench101、ResNet 为基准模型, 在 CIFAR-10 数据集上对比均匀稀疏性、全局稀疏性 2 种稀疏性策略。均匀稀疏性对网络所有层施加相同 PR, 而全局稀疏性根据重要性评分动态调整各层 PR, 但保证总体剪枝率与均匀稀疏性一致。本文中的稀疏性为通道稀疏性, 其中, 均匀稀疏性表示每一层的通道数缩减比例, 而全局稀疏性表示网络中的通道数总和在压缩后的缩减比例。比较不同稀疏性策略下模型在轻量化后的准确率, 表 5 给出了关于层稀疏性的实验结果。

表 5　不同分层裁剪策略的效果对比

Table 5　Comparison of the effectiveness of different hierarchical cutting strategies

网络架构	分组策略	压缩比例/%	识别准确率/%
ResNet56	全局稀疏	10	92.91
		30	91.05
		50	88.93
	均匀稀疏	10	92.59
		30	91.91
		50	90.84
NAS-Bench101	全局稀疏	10	95.81
		30	92.66
		50	91.05
	均匀稀疏	10	96.29
		30	96.10
		50	95.61

观察表 5 可以得出, 在大多数实验场景下, 全局稀疏性展现出了相较于均匀稀疏性更为出色的性能表现。通过全局统筹的剪枝策略, 能够更精准地识别并移除网络中真正冗余的连接和参数, 从而在有效减少模型计算量和存储需求的同时, 最大程度地维持模型的精度。在部分实验案例中, 本文也观察到其存在一定的局限性, 由于其全局化的决策机制, 有可能会对某些特定层进行过度剪枝, 这些层可能在模型的整体架构中承担着关键的功能, 尽管从全局角度看, 其部分连接或参数看似冗余, 但实际上对模型精度有着重要影响。一旦对这些关键层进行过度剪枝, 就会破坏模型的信息流传递, 导致模型对输入数据的特征提取和理解能力下降, 进而引发精度的显著下降。

3　结　论

1) 针对传统剪枝算法依赖人工经验、难以泛化至低空无人机机载设备中不规则神经网络架构的问题, 本文提出基于依赖图分析的结构化剪枝算法, 通过形式化建模网络参数间的拓扑依赖关系, 实现了复杂结构(如残差连接、跨层跳跃)的自动化分组与剪枝。实验表明, 在通用 ResNet 模型中, 本文算法相较人工策略更优, 且在 NAS-Bench101 不规则网络上, 可以在实现通用剪枝的同时, 保持较好的综合性能。

2) 本文方法的核心创新在于: 首先, 针对神经网络中可能存在的多分支、多类型依赖关系, 根据前向计算推导依赖图建立依赖传播机制, 支持多分支结构的耦合参数快速分组; 其次, 在参数组重要性判断时, 通过图同构判定, 实现异构子结构的可比性建模, 实现跨模块重要性的全局比较。

3) 在 NAS-Bench101 模型及 ResNet 系列网络中, 本文算法能够精准识别多分支连接、残差连接、拼接等复杂结构, 并顺利实施分组剪枝操作, 对网络实施 50% 的压缩比例进行轻量化处理时, 模型的识别准确率下降幅度控制在 1% 以内。实验结果表明, 本文提出的不规则结构剪枝算法具备通用性和有效性。

参考文献 (References)

[1] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 770-778.

[2] Wang H, Wang H, Liu Y J, et al. Real-time obstacle perception method for UAVs with an RGB-D camera in low-light environments[J]. Signal, Image and Video Processing, 2025, 19(3): 256.

[3] El Ghazouali S, Mhirit Y, Oukhrid A, et al. FusionVision: a comprehensive approach of 3D object reconstruction and segmentation from RGB-D cameras using YOLO and fast segment anything[J]. Sensors, 2024, 24(9): 2889.

[4] Saha A, Dhara B C. 3D LiDAR-based obstacle detection and tracking for autonomous navigation in dynamic environments[J]. International Journal of Intelligent Robotics and Applications, 2024, 8(1): 39-60.

[5] Wang C H, Wang Y C, Xu M Z, et al. Stepwise goal-driven networks for trajectory prediction[J]. IEEE Robotics and Automation Letters, 2022, 7(2): 2716-2723.

[6] Li R C, Katsigiannis S, Kim T K, et al. BP-SGCN: behavioral

pseudo-label informed sparse graph convolution network for pedestrian and heterogeneous trajectory prediction[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(8): 14566-14580.

[7] Xin G P, Chu D F, Lu L P, et al. Multiagent trajectory prediction with difficulty-guided feature enhancement network[J]. IEEE Robotics and Automation Letters, 2025, 10(3): 2582-2589.

[8] Cai Y J, Ge L H, Liu J, et al. Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks[C]// Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2020: 2272-2281.

[9] Cao Z, Simon T, Wei S H, et al. Realtime multi-person 2D pose estimation using part affinity fields[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 1302-1310.

[10] Mcnally W, Vats K, Wong A, et al. Rethinking keypoint representations: modeling keypoints and poses as objects for multi-person human pose estimation[C]//Proceedings of the Computer Vision-ECCV 2022. Berlin: Springer, 2022: 37-54.

[11] 姜晓勇, 李忠义, 黄朗月, 等. 神经网络剪枝技术研究综述[J]. 应用科学学报, 2022, 40(5): 838-849.
Jiang X Y, Li Z Y, Huang L Y, et al. Review of neural network pruning techniques[J]. Journal of Applied Sciences, 2022, 40(5): 838-849(in Chinese).

[12] Lecun Y, Denker J S, Solla S A. Optimal brain damage[J]. Advances in Neural Information Processing Systems, 1989, 2: 598-605.

[13] Han S, Mao H Z, Dally W J. Deep compression: compressing deep neural network with pruning, trained quantization and Huffman coding[EB/OL]. (2016-02-15)[2025-05-10]. https://arxiv.org/abs/1510.00149.

[14] Yang H C, Zhu Y H, Liu J. Energy-constrained compression for deep neural networks via weighted sparse projection and layer input masking[EB/OL]. (2019-06-02)[2025-05-10]. https://arxiv.org/abs/1806.04321.

[15] Gong R H, Yong Y, Wang Z N, et al. Fast and controllable post-training sparsity: learning optimal sparsity allocation with global constraint in minutes[C]//Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence. New York: ACM, 2024: 12190-12198.

[16] Li H, Kadav A, Durdanovic I, et al. Pruning filters for efficient ConvNets[EB/OL]. (2017-03-10)[2025-05-10]. https://arxiv.org/abs/1608.08710.

[17] 刘宇, 雷雪梅. 融合 MobileNetV3 特征的结构化剪枝方法[J]. 上海交通大学学报, 2023, 57(9): 1203-1213.
Liu Y, Lei X M. A structured pruning method integrating characteristics of MobileNetV3[J]. Journal of Shanghai Jiao Tong University, 2023, 57(9): 1203-1213(in Chinese).

[18] Hu H Y, Peng R, Tai Y W, et al. Network trimming: a data-driven neuron pruning approach towards efficient deep architectures[EB/OL]. (2016-07-12)[2025-05-10]. https://arxiv.org/abs/1607.03250.

[19] Liu Z, Li J G, Shen Z Q, et al. Learning efficient convolutional networks through network slimming[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 2755-2763.

[20] Lin M B, Cao L J, Zhang Y X, et al. Pruning networks with cross-layer ranking & k-reciprocal nearest filters[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(11): 9139-9148.

[21] Akimoto Y, Shirakawa S. Analysis of search space design for neural architecture search with weight sharing[C]//Proceedings of the 2024 International Joint Conference on Neural Networks. Piscataway: IEEE Press, 2024: 1-8.

[22] Wang H, Qin C, Bai Y, et al. Why is the state of neural network pruning so confusing on the fairness, comparison setup, and trainability in network pruning[EB/OL]. (2023-02-21)[2025-05-10]. https://arxiv.org/abs/2301.05219.

[23] Feng Q T, Xu K, Li Y H, et al. Edge-wise one-level global pruning on NAS generated networks[C]//Proceedings of the Pattern Recognition and Computer Vision. Berlin: Springer, 2021: 3-15.

[24] Jiang M N, Wang J C, Eldebiky A, et al. Class-aware pruning for efficient neural networks[C]//Proceedings of the 2024 Design, Automation & Test in Europe Conference & Exhibition. Piscataway: IEEE Press, 2024: 1-6.

[25] Chollet F. Xception: deep learning with depthwise separable convolutions[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 1800-1807.

[26] Chatzikonstantinou C, Konstantinidis D, Dimitropoulos K, et al. Recurrent neural network pruning using dynamical systems and iterative fine-tuning[J]. Neural Networks, 2021, 143: 475-488.

[27] Wei S Y, Ye T Z, Zhang S, et al. Joint token pruning and squeezing towards more aggressive compression of vision Transformers[C]// Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 2092-2101.

[28] Iandola F N, Han S, Moskewicz M W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size[EB/OL]. (2016-11-04)[2025-05-10]. https://arxiv.org/abs/1602.07360.

Low-altitude perception models pruning algorithm with irregular structures

YANG Zi，ZHUANG Liansheng*

(School of Cyber Science and Technology，University of Science and Technology of China，Hefei 230026，China)

Abstract：Large neural networks are challenging to implement on terminal devices due to restricted processing resources, notwithstanding deep learning's superior performance in low-altitude intelligent sensing. Pruning techniques reduce model complexity by eliminating redundant parameters, thereby improving real-time performance. However, existing methods primarily target regularly structured models and struggle to handle irregular structures generated by neural architecture search (NAS). To address this, a dependency graph-based pruning scheme for irregular structures is proposed to optimize model efficiency. The method first employs structural parsing techniques of computational graphs to automatically identify unique multi-level connection patterns in low-altitude perception models. Subsequently, it constructs an efficient parameter grouping module to precisely locate minimum functional units that can be completely removed. Finally, a cross-layer comparison global optimization strategy is adopted to implement unified evaluation and pruning of isomorphic substructures. This solution overcomes the technical constraints of structural adaptability when compared to current approaches, allowing global precision compression for irregular network architectures. This feature makes it especially appropriate for low-altitude perception scenarios that call for intricate spatial relationship processing. Experimental results demonstrate that the proposed irregular structure pruning method can effectively identify and perform group pruning on complex structures, including multi-branch connections, residual connections, and concatenations, in both NAS-Bench101 models and ResNet series networks. When implementing 50% compression rate network lightweighting on the CIFAR-10 dataset, the recognition accuracy drop does not exceed 1%.

Keywords：low-altitude perception；convolutional networks；lightweight models；model pruning；irregular network architectures

Received：2025-05-19；**Accepted**：2025-05-30；**Published Online**：2025-06-05 14：26

URL：link.cnki.net/urlid/11.2625.V.20250605.1115.003

* **Corresponding author**. E-mail：yangzi@mail.ustc.edu.cn