

DOI: 10.13382/j.jemi.B2508622

基于 CNN-LRP 层级适应性冻结的模型优化方法

王振宇 赵科东 孙永荣 杨杰 陈旗

(南京航空航天大学导航研究中心 南京 211106)

摘要:针对外观多样且样本稀缺的信号灯检测场景,迁移学习已成为主流解决方案。然而,传统全参数微调和浅层冻结未量化网络层对任务具体贡献,训练更新范围过广、参数冗余,在小样本条件下易过拟合从而限制模型性能。为此提出一种基于卷积神经网络-分层相关性传播(CNN-LRP)层级适应性冻结的模型优化方法。基于正向相关比、空间聚焦度与层深惩罚因子,将CNN-LRP的像素级归因拓展至卷积层级,构建层级贡献度评估体系,实现解释性分析与冻结策略的融合。针对传统策略任务适应性不足的问题,设计了多指标融合的卷积层适应性评分方法,自动生成冻结策略,通过冻结低适应性层,实现任务驱动的迁移,从而优化模型。最后,利用不同光照与距离的场景数据进行验证测试。结果表明,相比于全微调与其他冻结策略,该方法生成的模型在测试准确率上提升达3.1%,参数调优量下降34.49%,在精度、效率、泛化性、鲁棒性与卷积网络普适性等方面具有更好的效果。

关键词: 迁移学习; 小样本学习; 域适应; 信号灯状态监测; 模型微调; 模型优化

中图分类号: TP391.41 **文献标识码:** A **国家标准学科分类代码:** 520.6040

LATS: Model optimisation method based on CNN-LRP layer-wise adaptability freezing

Wang Zhenyu Zhao Kedong Sun Yongrong Yang Jie Chen Qi

(Navigation Research Center, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

Abstract: In multi-scenario perception and intelligent control systems, the accurate and efficient recognition of traffic light states is an important support for ensuring stable system operation and the collaborative operation of intelligent devices. In response to the diverse appearances and scarce samples in traffic light detection scenarios, transfer learning has become a mainstream solution. However, traditional full parameter fine-tuning and shallow freezing do not quantify the specific contributions of network layers to the task, leading to overly broad training updates, parameter redundancy, and a tendency to overfit under small sample conditions, which limits model performance. To address this, this paper proposes a layer-adaptive guided freezing strategy generation method based on CNN-LRP. By utilising the forward correlation ratio, spatial focus degree, and layer depth penalty factor, the pixel-level correlation of CNN-LRP is expanded to the convolutional layer level, constructing a layer contribution evaluation system that integrates interpretative analysis with the freezing strategy. To tackle the issue of insufficient task adaptability in traditional strategies, this paper designs a multi-metric fusion convolutional layer adaptability scoring method that automatically generates freezing strategies, and realises task-driven migration by freezing the low-adaptability layer, so as to optimise the model. Finally, verification tests were conducted using scene data with different lighting and distances. The results show that compared with full fine-tuning and other freezing strategies, the model generated by the method proposed in this paper has an improvement of 3.1% in test accuracy, a reduction of 34.49% in parameter tuning volume, and better performance in terms of accuracy, efficiency, generalization, robustness, and the universality of convolutional networks.

Keywords: transfer learning; model fine-tuning; small sample learning; domain adaptation; traffic light status monitoring; model optimisation

0 引言

在工业自动化^[1]、海上平台监控^[2]、空中加油^[3]等领域,信号灯作为关键状态指示工具,其识别准确性对保障人机协同与自主决策至关重要。

近年来,深度学习在信号灯状态识别等视觉感知任务中得到广泛应用^[4-6],但其对大规模标注数据的依赖时期难以适应高风险或涉密场景,易出现特征学习不足与过拟合现象,影响识别精度与泛化性^[7]。

为缓解小样本条件下的训练瓶颈,迁移学习被广泛用于提升模型的泛化能力^[8],其中,冻结策略的设置对模型在目标任务中的性能具有关键影响^[9]。Dar 等^[10]指出,冻结层数与其位置直接影响迁移效果,冻结不当易导致源域泛化与目标域适应能力同步下降,显著削弱模型性能。

深度神经网络通常被认为浅层提取边缘等低级特征,深层捕捉语义等高级表征。然而,Yosinski 等^[11]发现,中间层同时具备通用性与任务特定性,盲目冻结可能破坏层间协同,降低模型适应能力。但传统冻结策略多基于经验划分层级,忽略了中间层在任务迁移中的独特价值,限制了迁移性能的进一步提升。

针对上述问题,部分研究尝试引入新的冻结策略。Chen 等^[12]针对多任务网络提出梯度归一化算法,在提升多任务准确率的同时降低了过拟合风险;Brock 等^[13]提出逐层冻结策略,在保证准确率基本稳定的前提下,显著减少了训练时间;Sorrenti 等^[14]在训练前测试多种冻结掩码并自动选取最优组合,实现了任务适应性与旧任务记忆能力之间的平衡。

尽管上述方法在一定程度上缓解了冻结策略的局限性,但仍主要依赖经验或测试验证过程,缺乏对网络层级任务相关性的量化评估,导致策略生成过程仍具有显著的不确定性与主观性。因此,如何揭示模型各层与目标任务的相关性,是构建更加精确、稳健的冻结策略的核心挑战。

可解释性方法为任务相关性量化提供了非侵入式、数据驱动的路径^[15-16]。与基于梯度的解释方法不同,分层相关性传播^[17](layer-wise relevance propagation, LRP)通过逐层反向传播“相关性分数”,显式量化输入特征对模型预测的贡献。其层级相关性守恒机制利用 z-rule、 ϵ -rule、 $\alpha\beta$ -rule 等传播规则^[18]有效避免了梯度饱和与消失引发的解释失真,具备良好的结构感知能力与稳定性。

为克服传统 LRP 方法仅限于全连接网络,难以适用于卷积神经网络(convolutional neural network, CNN)中权重共享与降采样结构的局限,Zang 等^[19]提出基于卷积神经网络的分层相关性传播方法(CNN-LRP),通过融合局

部连接权重与激活强度,构建输出至各卷积层的归因通路,为层级贡献分析提供了理论依据与实现路径。

然而,当前将可解释性方法引入冻结策略的研究尚处于初步阶段,仍需解决如下关键问题。

1) 现有 CNN-LRP 方法主要聚焦于输入像素级的贡献度分析,忽略了中间层级在任务判别中的作用,难以为冻结策略提供结构层面的量化依据。

2) 多数冻结策略依赖经验,缺乏对不同层与目标任务关联性的刻画,导致可解释性与适应性不足,迁移性能易受波动。

3) 部分方法需对模型结构进行修改或依赖测试验证过程,工程适配性差,难以满足小样本场景中对稳定性与简洁性的需求。

为此,本文提出一种基于 CNN-LRP 层级适应性冻结的模型优化方法(layer-wise adaptability-guided transfer strategy, LATS)。该方法借助模型可解释性中的归因分析机制,在像素级归因的基础上对卷积神经网络的各层进行任务贡献度的量化评估,识别出对目标任务具有关键性贡献的结构单元,通过冻结适应性低的冗余卷积层,引导模型聚焦于高相关性特征的学习,显著增强模型的判别能力与泛化性能。相比于传统经验驱动的全微调冻结方式,可提升模型的可迁移性与鲁棒性。

1 基于 CNN-LRP 的层适应性冻结策略框架

1.1 信号灯视觉特性与跨域可迁移性分析

在实际应用中,由于采集成本高昂以及信息保密等限制,吊舱信号灯、机床信号灯等特定场景下的目标样本往往数量有限,难以支撑从零开始的模型训练。然而,信号灯目标在任务设定及视觉表达上与交通信号灯等公共数据存在较强的相似性,为跨领域迁移学习提供了可能。

在任务层面,信号灯状态的判别主要依赖于两项核心视觉特征:1) 颜色通道的差异性,不同状态(红、黄、绿)在色彩分布上区分明显,是状态识别的首要依据;2) 空间排列的规则性,信号灯常采用垂直或水平的固定布局,使不同颜色间呈现稳定的空间位置关系,增强了结构特征的可迁移性,也为模型学习提供了先验支持。

交通信号灯与吊舱信号灯在 RGB 颜色均值特征上的 t-SNE^[20]降维结果如图 1 所示。可以看出绿色类别在源域与目标域之间形成了显著的聚类趋势,表明该类别在通道响应上具有较高的一致性,为领域迁移提供了直观支持。相比之下,红色与黄色类别则表现出明显的特征分布偏移,揭示了跨域迁移中的域间差异,需要采用任务引导迁移策略提高模型目标域适应能力与泛化能力。

1.2 层级适应性引导冻结策略结构框架

为明确本文方法在整体模型训练流程中的位置,本

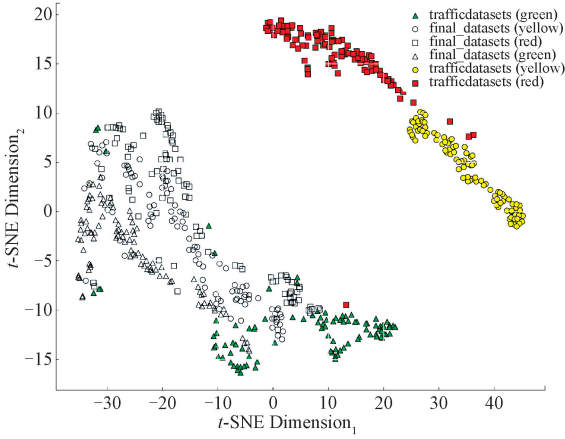


图1 交通信号灯与吊舱信号灯的 t-SNE 降维结果

Fig. 1 Results of t-SNE downscaling for traffic signals and pod signals

文将模型训练分为模型预训练阶段、冻结策略生成阶段和模型微调阶段3个部分。

针对当前冻结策略依赖经验、缺乏任务适应性支撑导致模型泛化性不足的问题,在冻结策略生成阶段将CNN-LRP的归因机制由传统的输入空间拓展至网络结构空间,通过分析中间层级的相关性分布,量化不同网络层对目标任务的贡献程度;随后,结合多输入样本下的层级贡献度的平均分数及稳定性,构建任务相关的层适应性评分函数,评估各层的重要性,生成层级任务适应性引导冻结策略;最后提出冻结策略生成模块的轻量集成方法,在不干扰模型参数的前提下自动筛选出对目标任务适应性较低的冗余层进行冻结。该方法整体结构框图如图2所示。

2 面向任务的层贡献度建模与冻结评估机制设计

2.1 层级多维贡献度聚合方法

传统LRP方法聚焦于元素级的归因,能够呈现较强的可视化结果,但难以作用于网络结构级别的解释。

为实现以“层”为单位的结构冻结与微调,本文提出一种层级多维贡献度聚合方法,用于将神经元级的归因结果聚合为层级评分,形成层级贡献度的可比较指标,从而为迁移学习过程中的冻结策略提供依据。

首先需对其神经元相关性进行空间归约处理,从而衡量每一层在整个模型决策中的结构性贡献。本文借鉴了CNN-LRP的相关定义,其相关性传播可表示为:

$$z[n, c, i, j, u, v] = a^{(l)}[c, i + u, j + v] \cdot w^{(l+1)}[n, c, u, v] \quad (1)$$

式中: n 表示输出通道编号; c 为输入通道编号; i, j 表示

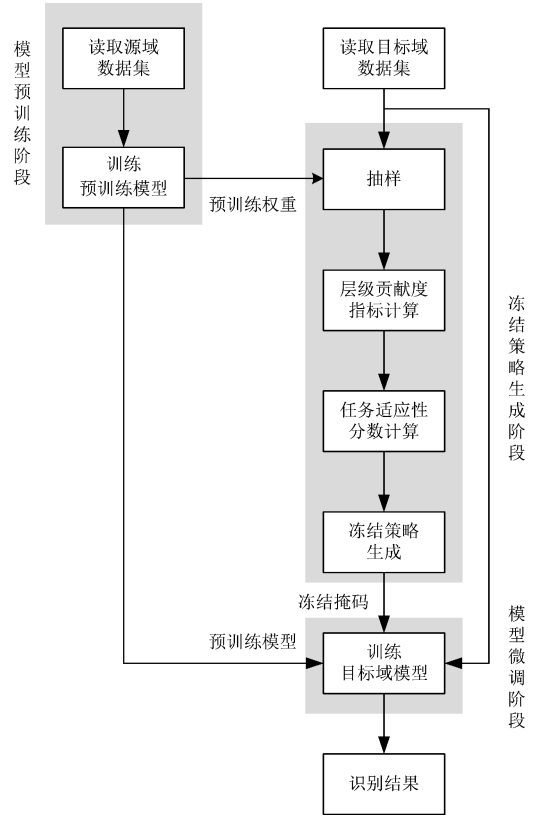


图2 冻结策略生成结构框图

Fig. 2 Diagram of freezing strategy generation

卷积滑动窗口当前在输出特征图上的位置; u, v 为卷积核内部的相对偏移坐标,取值范围为 $0 \leq u, v < k$, 其中 k 为卷积核空间尺寸; $z[n, c, i, j, u, v]$ 为输入特征图第 c 个通道中,位置 $(i + u, j + v)$ 的激活值与卷积核权重在该位置的乘积,表示该输入点对输出特征图第 n 个通道位置 (i, j) 的局部贡献度; $a^{(l)}[c, i + u, j + v]$ 为第 l 层输出(即第 $l + 1$ 层输入)中,第 c 个通道在空间位置 $(i + u, j + v)$ 的激活值; $w^{(l+1)}[n, c, u, v]$ 为第 $l + 1$ 层中,第 n 个卷积核,在第 c 个输入通道下,卷积核内部位置 (u, v) 的权重参数。

则 l 层输入特征图中第 c 个通道中 $(i + u, j + v)$ 处的像素相关性可改写为:

$$r^{(l)}[c, i + u, j + v] = \sum_n \frac{z[n, c, i, j, u, v]}{\sum_{c'} \sum_{u', v'} z[n, c', i, j, u', v'] + \varepsilon} \times r^{(l+1)}[n, i, j] \quad (2)$$

式中: $r^{(l+1)}[n, i, j]$ 为输出特征图第 n 个通道中 (i, j) 像素的相关性; $\sum_{c'} \sum_{u', v'} z[n, c', i, j, u', v']$ 表示输出特征图第 n 个通道中 (i, j) 像素的所有输入贡献和; ε 为防止分母为0的微小正数。

由式(1)、(2)可知,CNN-LRP反向传播生成的每一层输出相关性张量 $r^{(l)} \in r^{C \times H \times W}$ 具有通道结构,不同通道

表征不同语义方向的特征响应,如图 3(a)所示。为了获得该层整体的空间相关性分布,通过通道维度的求和,如图 3(b)所示,得到大小为 $H \cdot W$ 的二维相关性特征图中位于 (x, y) 位置的像素相关性:

$$R^{(l)}(x, y) = \sum_{c=1}^C r^{(l)}[c, x, y] \quad (3)$$

式中: c 为通道编号; C 为总通道数。

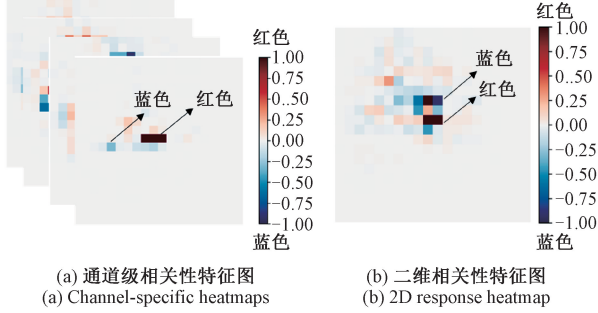


图3 相关性特征图的通道维度展开与层级归因结果

Fig. 3 Channel dimension expansion and hierarchical attribution results for correlation feature maps

图3为当前层在空间维度上的总体相关性强度分布,红色区域表示该层在该位置有较强的正向激活相关性,蓝色则为负向相关性,作为后续构建层级评分指标的基础。

2.2 多维度层贡献度指标设计

为量化层级贡献度,结合二维相关性特征图,构建层贡献度指标,将归因结果从像素级向网络结构空间拓展,根据任务特点,量化中间层在预测决策中的重要性。

1) 正向相关比

任务相关性有正相关以及负相关性,正相关性代表该单元对预测结果有正向作用,负相关性代表该单元存在冗余或抑制性激活,对模型泛化构成干扰,因此设计正向相关比:

$$\alpha^{(l)} = \frac{\sum_{R^{(l)}(x, y) > 0} R^{(l)}(x, y)}{\sum_{x, y} |R^{(l)}(x, y)|} \quad (4)$$

如果该层产生大量正相关像素 ($r^{(l)} > 0$),说明该层有利于适应目标域任务。通过衡量二维相关性特征图中正相关像素的占比,反映该层对模型当前预测结果的有利性,比值越大则对任务越有贡献。在冻结策略中,结合正向相关比指标可实现对高干扰层的识别与剔除。

2) 空间聚焦度

LRP 本身满足相关性守恒原则,即某一层的正负相关性总和与上一层保持一致,但相关性在每层内的空间分布不同。在信号灯状态识别任务中,仅有图像中局部区域(例如吊舱信号灯区域)对分类起到决定性作用,因此若某一层的输出特征在相关性分布上呈现高度聚焦于

关键区域,说明该层在任务判别中具有更高的贡献度。

结合 Top-K 占比指标衡量像素相关性在空间分布上的集中性:

$$\beta^{(l)} = \frac{\sum_{\text{Top-}k\%} R^{(l)}(x, y)}{\sum_{x, y} |R^{(l)}(x, y)|} \quad (5)$$

本文采用 Top-10% 相关性区域占比指标,以衡量预测所依赖的相关性是否集中于图像中少量关键区域。该指标通过计算二维相关性特征图中得分排名前 10% 的区域所占总相关性比例,反映模型是否能够聚焦于有限而关键的判别区域,而非对整张图像产生广泛且分散的响应。通过筛选有效学习到目标任务中的核心语义特征的层,冻结依赖于冗余背景或噪声信息进行判别的层,提高模型的判别性与泛化能力。

3) 层深惩罚因子

由于卷积神经网络在层级结构中存在“浅层侧重局部低级特征、深层聚焦抽象语义特征”的特性,导致深层通常呈现出更集中的高相关性区域。为避免这种“后层天然高分”现象对贡献度评估产生偏差,本文引入归一化层深惩罚因子 $\gamma^{(l)}$ 作为惩罚项,对各层的原始贡献得分进行层级补偿:

$$\gamma^{(l)} = 1 - \lambda \cdot \frac{l}{L} \quad (6)$$

式中: l 为当前卷积层编号; L 为总层数; $\lambda \in [0, 1]$ 为惩罚系数,默认为 0.2~0.5。通过将每一层的归一化深度作为贡献度得分的调制系数,抑制因网络深度带来的评分偏置,确保各层贡献的可比性与公平性。

最终层贡献度得分 $S^{(l)}$ 由 3 个变量组成:

$$S^{(l)} = \alpha^{(l)} \cdot \beta^{(l)} \cdot \gamma^{(l)} \quad (7)$$

式中: $\alpha^{(l)}$ 保证语义相关性; $\beta^{(l)}$ 保证空间聚焦; $\gamma^{(l)}$ 避免语义偏移。这 3 个变量共同构成具有可比性与可解释性的层选择标准。该层级评分机制不依赖具体任务标签或领域先验,仅基于当前预测结果的反向信号,可在迁移学习阶段对任意任务进行结构解释,为后续的冻结决策提供量化依据。

2.3 任务适应性评分函数构建及冻结策略生成

在层级归因机制得出的各层贡献度的基础上,本文通过分析层贡献的显著性特征与分布稳定性,构建任务适应性评分函数,进而制定层级冻结策略,以提升模型迁移学习的效率与泛化性能,实现有效知识的精准迁移。

式(7)综合考虑了该层产生的正相关性强度、空间聚焦度以及层深惩罚因子,仅对其进行排序筛选无法衡量模型在不同输入图片上的稳定性,因此本文在单图像贡献打分的基础上,引入统计稳定性评估指标:

$$\mu_l = \frac{1}{N} \sum_{i=1}^N S_i^{(l)} \quad (8)$$

$$\sigma_l = \sqrt{\frac{1}{N} \sum_{i=1}^N (S_l^{(i)} - \mu_l)^2}$$

(9)

$$CV_l = \frac{\sigma_l}{\mu_l + \varepsilon}$$

(10)

式中: N 为总样本数; $S_l^{(i)}$ 为 l 层在对 i 张图片识别结果的贡献度; μ_l 为 l 层在不同样本图像中的平均贡献度, 反映该层对最终模型输出的重要性; σ_l 为 l 层在不同样本图像中的贡献度标准差; CV_l 为变异系数, 衡量 l 层重要性的稳定性; ε 为防止分母为 0 的微小正数。

通过计算每层在多个输入样本上的平均贡献度 μ_l 及其稳定性指标(变异系数 CV_l), 设计一种多指标融合的层适应性评分计算方法:

$$Score_l = \mu_l \cdot (1 - \xi \cdot \min(CV_l, 1))$$

(11)

式中: $\xi \in [0, 1]$ 为稳定性权重参数, 默认为 0.7, 为防止变异系数过大导致得分畸变, 采用截断操作 $\min(CV_l, 1)$ 。通过该层适应性评分函数, 保留高贡献度且稳定的层(不冻结), 优先冻结贡献度波动大或平均贡献度较低的层。

通过对所有层的 $Score_l$ 进行排序, 可选取得分最低的前 K 层作为候选冻结层, 构成最终的优化冻结策略。

2.4 冻结策略生成模块的轻量集成方法

为实现结构层级的冻结策略, 本文设计了一种非侵入式、可兼容的结构归因模块。该模块通过对源域预训练模型在目标域样本上的响应进行分析, 评估各卷积层对输出的层级贡献度, 在不干扰原始模型结构与参数的前提下, 自动生成冻结策略, 引导迁移。

模块首先加载交通信号灯检测任务中训练得到的预训练模型, 并将其设为推理模式(model.eval()), 避免批归一化(batch normalization, BN)与随机失活(dropout)层在目标域图像上的不稳定行为引入扰动。为保证梯度反向传播路径的完整性, 进一步关闭非极大值抑制(non-maximum suppression, NMS)过程, 保留全部候选框; 同时对模型中的所有操作显式禁用所有原地操作(in-place), 避免梯度冲突或计算图破坏, 从而保证梯度信息的稳定性与可导性。

接着, 在每一个具有可学习参数的卷积层上注册前向与反向钩子函数, 分别记录激活输出与反向梯度, 实现层级贡献度估计。在推理过程中, 该模块使用已标注的目标域图像(如 30 张)进行逐图归因, 基于正向相关比、空间聚焦度及层深惩罚因子, 构建层级贡献度打分函数。在此基础上, 通过统计多个图像样本中各层贡献度的均值、标准差与变异系数, 评估各层贡献的稳定性与重要性, 最终生成冻结掩码。该掩码用于在后续训练中冻结结构适应性较低的卷积层, 仅保留对当前任务响应显著的关键层级, 从而提高模型在迁移学习过程中的效率与适应性。

整个结构归因分析流程可在迁移训练前独立完成, 既不依赖额外监督信号, 也不影响主模型参数或结构, 具备良好的模型通用性、轻量性与可解释性。同时, 该模块可拓展应用于剪枝、知识蒸馏等结构优化任务, 展现出较强的通用性。

3 实验设计

为验证本文提出方法的泛化性、准确性、鲁棒性及卷积网络普适性, 设计验证实验, 以空中加油吊舱信号灯状态识别为典型指示灯场景, 对本文方法展开验证。

3.1 实验平台搭建与数据集获取

在室外环境下, 利用信号灯、吊舱模型、可见光相机、STM32 开发板, 385 模组相机、移动电源和推车等设备设计并搭建吊舱信号灯状态识别地面实验平台, 在 3 种不同环境下采集图像数据模拟空中加油对接过程中信号灯状态, 搭建平台如图 4 所示。

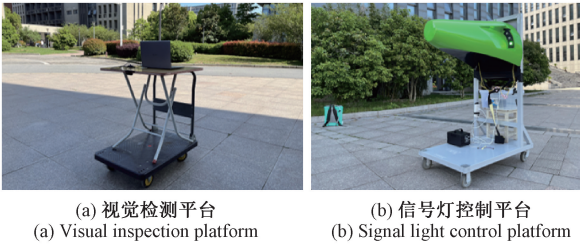


图 4 地面模拟信号灯识别实验平台

Fig. 4 Ground simulation signal lamp recognition experimental platform

本文用网络模型训练的硬件平台及用于测试的硬件平台各项设备的具体参数如表 1 所示。

表 1 计算机设备参数信息

Table 1 Computer equipment parameter information

名称	CPU 型号	内存/GB	GPU 型号	显存/GB
深度学习机	Intel Xeon	64	Nvidia	24
	W-3175X		GTX3090	
桌面端电脑	Intel Core	16	Nvidia	6
	i7-10700F		GTX1660s	

试验训练所用操作系统为 Ubuntu20.04, 检测所用操作系统为 Windows 11。深度网络模型的训练和检测条件, 系统环境具体配置如表 2 所示。

表 2 环境配置信息

Table 2 Environment configuration information

名称	网络框架	库版本号	CUDA 版本	cuDNN 版本	Python 版本	OpenCV 版本
训练	Pytorch	1.13.0	11.7	8.7.0	3.8.0	4.5.5
检测	Pytorch	1.13.0	11.3	8.9.3	3.8.0	4.0.0

实验所用的数据集包括网络搜集的交通信号灯数据集,以及地面模拟采集的吊舱信号灯数据集。

考虑到空中加油场景的特殊性,目前尚无公开可用的标准数据集,因此利用迁移学习的特性,构建了与吊舱信号灯特征同分布的交通信号灯样本数据库,其中包含 31 178 张图片,如图 5 所示。



图 5 部分交通信号灯数据集

Fig. 5 Partial traffic signal light dataset

同时从检测模型的训练需求出发,采集 5~20 m 处图像,构建专用于空中加油吊舱信号灯状态识别的自定义小样本数据库。所使用的数据集包 300 张图片,每种状态 75 张,由地面试验平台模拟采集而来,如图 6 所示。其中,地面模拟对接平台中所使用的空中加油吊舱信号灯是依据真实信号灯制作的等比例模型。



图 6 部分吊舱信号灯数据集

Fig. 6 Partial pod signal lamp dataset

3.2 冻结策略生成实验分析

1) 冻结策略参数消融实验

为衡量惩罚系数 λ 及稳定性权重参数 ξ 对冻结策略的影响,本文设计了双参数消融实验,由式(6)和(11)可知,二者在数学上无交叉项,具有独立性,因此可分别固定另一参数进行单变量搜索以获得联合最优组合。惩罚系数及稳定性权重对验证准确率的影响如图 7 所示。

惩罚系数 λ 用于平衡细节与语义的重要性, λ 过小会导致深层特征权重过大、语义层冻结不足,而 λ 过大则会过度压制语义层信息。实验中 $\lambda = 0.4$ 时模型性能最优,体现了对细节与语义信息的合理平衡。

稳定性权重 ξ 用于惩罚波动大的不稳定层,抑制扰动,适度提升 ξ 可削弱扰动、增强冻结层选择的稳定性,但过高会导致过度平滑。实验结果表明当 $\lambda = 0.4, \xi = 0.7$ 时测试精度最高(81.1%),表明适度的层深惩罚与

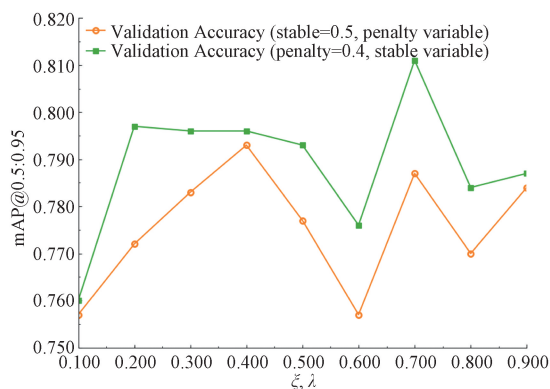


图 7 稳定性权重与惩罚系数对准确率的影响

Fig. 7 The impact of stability weight and penalty coefficient on accuracy

稳定性约束可使冻结策略更准确地识别结构稳定层与任务相关层,实现更合理的冻结决策。

2) 冻结策略生成

首先通过 CNN-LRP 对 YOLOv5 网络中所有卷积层进行了归因分析。采用 CNN-LRP 传播规则,结合层级多维贡献度聚合方法得到二维相关性特征图,在此基础上结合正向归因比、空间聚焦度与层深惩罚因子,构建综合贡献度评分规则,计算每层在特定检测任务中的贡献度。通过在多张输入图像上评估层贡献度的稳定性,计算层贡献度及稳定性的层任务适应性得分,生成冻结策略。图 8 所示为各卷积层的归一化得分分布情况,并据此设置冻结阈值,实现可解释性驱动的分层冻结策略。

从图 8 可以观察到,网络前部的浅层卷积层的平均得分较高,意味着其在目标检测任务中贡献度高且对于不同图像稳定性较强,与低层提取通用特征的理解相符合,在第 6 层开始出现低适应性卷积层,这为结构化冻结策略提供了数据支持,相较于全模型微调,在训练资源有限或小样本场景下,通过冻结低适应性层,可在保留模型表达能力的同时,减少训练计算量,体现了本文冻结策略方法的可解释性及任务适应性指导能力。

3) 冻结策略生成时间评估

本文方法在推理阶段基于前向传播引入 CNN-LRP 反向归因机制,并结合任务适应性评分以生成冻结策略。为量化 CNN-LRP 反向归因的引入对计算成本的影响,评估所提策略生成方法的时间开销,本研究对比了未引入 CNN-LRP 的原网络与应用该方法的网络在相同样本集上的平均推理耗时,在 Nvidia GTX3090 环境下,实验耗时如表 3 所示。

实验结果表明,本文方法相对前向传播时间开销增加了约 3.8 倍。经分析发现,额外开销主要源于 3 部分: 1) CNN-LRP 依赖反向传播计算梯度,而常规反向传播耗时通常是前向的 2~3 倍; 2) CNN-LRP 需逐元素操作并

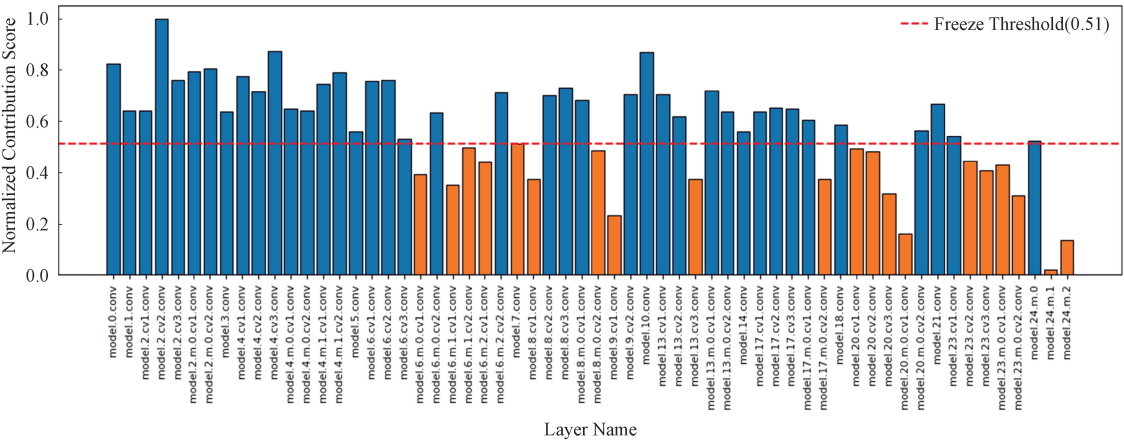


图 8 YOLOv5 各层任务适应性得分图

Fig. 8 Task adaptation score map for each layer of YOLOv5

逐层传播相关性,其计算量通常高于简单的反向传播;
3)策略生成阶段需额外计算层间任务适应性评分,进一步增加了耗时。3 者叠加导致额外开销高于纯反向传播。

表 3 冻结策略推理耗时对比
Table 3 Comparison of freezing strategy time consumption

模式	每帧平均耗时/ms	相对开销
原网络	21.7	—
LATS	103.9	+378.8%

尽管 CNN-LRP 增加了计算成本,但策略生成阶段通常为离线预处理步骤,不会影响模型在线推理的实时性。因此,该额外开销在实际应用中是可接受的。

3.3 冻结策略对比分析验证

为验证层级适应性引导冻结策略的有效性,在统一使用本文所述训练样本和同一台设备上(CPU 为 Intel Xeon W-3175X;GPU 为 GTX3090;RAM:64GB)上,开展冻结策略方法对比实验,评估本文基于 CNN-LRP 的层适应性引导冻结策略(LATS)、全微调冻结策略(Full)、标准冻结策略(Freeze 3/8/backbone)、逐层冻结策略(Freeze out)^[13]、GradNorm 冻结策略^[12]在验证样本上的分类效果。其中标准冻结策略分为冻结前 3 层、冻结前 8 层及冻结 backbone,常用的微调分类层方法因精度过低不予展示。其中冻结 Backbone 为冻结 model.0~model.9, FreezeOut 通过预设冻结计划,线性冻结前 9 层,分别在 0、43、86、130、173、217、260、304、347 epoch 将下一层加入冻结,同时保持原先冻结的层不变。GradNorm 方法结合梯度范数,动态调整每层的损失权重,使重要层(如检测头部)获得更大更新幅度,低重要层(如早期卷积层)更

新幅度减小,达到近似冻结效果,设置 GradNorm 强度 alpha 为 0.5 以保持层间平衡。

为进一步验证冻结策略的有效性,表 4 为不同冻结策略下的实验对比结果,均使用早停策略。其中参数调优量是当前冻结策略下仍参与梯度更新的参数数量与模型总参数量的比值,能够反映不同冻结策略下模型的训练复杂度。

表 4 实验结果
Table 4 Experimental results

冻结策略	mAP@ 0.5:0.95	参数调优量/%	Speed epoch/s
Full	0.780	100	3.026
Freeze 3	0.779	99.42	3.084
Freeze 8	0.727	66.83	3.008
Freeze backbone	0.713	40.67	2.813
Freezeout	0.776	50	3.067
GradNorm	0.773	100	3.258
LATS	0.811(↑)	65.51(↓)	3.001(↓)

由表 4 可知,LATS 方法在测试平均精度均值(mAP) mAP@ 0.5:0.95 上取得了 0.811 的最优性能,相较于 Full 提升了 3.1%,相比标准冻结策略中冻结前 3、8 层和 backbone 分别高出了 3.2%、8.4%和 9.8%,相比 Freeze out 的逐层冻结策略高出了 3.5%,相比 GradNorm 冻结策略高出了 3.8%。

图 9 所示为各阶段检测模型的综合性能,气泡大小对应参数调优量,实线边缘的气泡代表本文方法,虚线边缘的气泡代表其他方法。

不同冻结策略具有局限性,结合表 4 可知,Full 参数调优量高达 100%,相较于标准冻结策略得到了充分的训练,取得了较高的精度(0.780),但因需要对全部参数进行更新,训练开销最大。相比之下,标准冻结策略(如

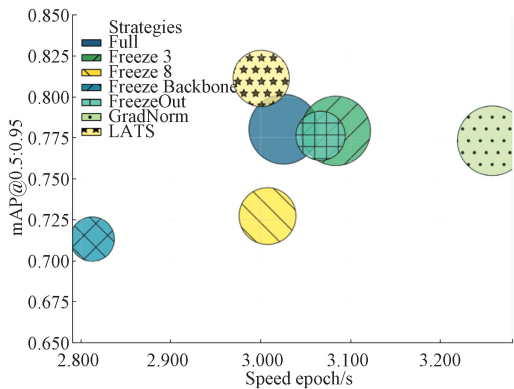


图9 各策略性能可视化对比

Fig. 9 Comparative visualisation of the performance of various strategies

Freeze 8、Freeze backbone)降低了参数调优量(分别至66.83%、40.67%),在计算效率上有所提升,但由于依赖经验忽略了层级与目标任务的关系,导致测试精度显著下降(最低降至0.713),表明网络未学习到关键知识,盲目冻结限制了模型的学习能力。而如Freezeout和GradNorm等自适应策略,虽具数据驱动性,但缺乏任务语义指导,导致模型精度受限(0.776和0.773)。

相比之下,本文方法通过冻结任务低适应性的卷积层,减少了冗余训练参数,使有限的样本量更有效地用于高相关层的参数更新,提高了准确率(0.811)。由表4可知,在冻结低适应性层后,本文方法相比于全微调冻结策略参数调优量降至65.51%(较全微调降低34.49%),单轮训练时间略减至3.001 s。由图9可知,本文方法在提升性能的同时提高了训练效率,实现了高精度与计算效率之间的平衡。综上本文方法在准确率上相较其他冻结策略有显著提升,同时在训练效率方面也具备优势,验证了本文冻结策略方法的有效性。

3.4 策略泛化性验证实验

为验证本文冻结策略的泛化性,在统一使用本文所述训练样本的基础上,开展不同场景对比实验,评估本文方法在验证样本上的信号灯状态识别准确率效果,结果如表5所示。本文设计了3种模拟远近切换动态环境的测试场景,其中场景1为强曝光条件下(距离跨度为20~30 m),场景2为阴天光照条件下(距离跨度为20~30 m),场景3为阴天光照条件下(距离跨度为20~40 m)。

表5 泛化性性能			
Table 5 Generalisation performance (%)			
冻结策略	场景1 识别准确率	场景2 识别准确率	场景3 识别准确率
Full	37.2	97.3	58.9
LATS	43.0(+5.8)	99.3(+2)	61.2(+2.3)

由表5可知,远距与强曝光环境显著降低模型识别精度。相较于Full,LATS方法通过冻结低适应性卷积层,抑制背景区域干扰,有效提升了模型对关键目标特征的响应能力,在3个典型泛化场景中均提升了识别准确率,说明模型学到了更具有针对性的特征,验证了本文方法提高模型泛化性的能力。

3.5 策略鲁棒性实验

为验证本文所提出的模型优化方法在不同干扰条件下的鲁棒性,本文在测试集中引入高斯噪声与随机遮挡两种典型干扰,对比原始YOLOv5模型与采用本文方法的YOLOv5模型的性能变化,实验结果如表6所示。

表6 鲁棒性实验			
Table 6 Robustness experiment (%)			
冻结策略	原始数据准确率	高斯噪声准确率	随即遮挡准确率
Full	78.0	76.8	76.8
LATS	81.1(↑)	78.0(↑)	78.6(↑)

实验表明,与全参数可训练的基准模型Full相比,LATS方法在3种条件下均保持更高的准确率,其中原始数据准确率提升3.1%,在高斯噪声与随机遮挡干扰下分别提升1.2%与1.8%。

这一现象表明,LATS通过冻结低适应性的层,使训练过程能够聚焦于具有稳定梯度和显著判别性的层,从而提升了特征学习的有效性。相比于全参数可训练的模型,LATS生成的冻结策略能够引导模型在训练阶段形成更一致、更具鲁棒性的特征空间,使得模型在输入分布变化或干扰条件下仍能保持稳定的判别能力,证明了本文方法的鲁棒性。

3.6 卷积网络普适性实验

本文方法基于CNN-LRP的层相关性传播机制,式(1)中的参数仅描述前向传播和反向传播之间的数学关系,其本质是通过层间的线性映射规则,计算神经网络中输入单元对输出的贡献,因此当模型具有显式分层结构,能够通过反向传播获取梯度,且具有相应传播规则,则可计算贡献度并生成冻结策略。

为验证方法的普适性,本文针对YOLOv5、Faster-RCNN及ResNet50卷积网络模型进行适应性改进,其中结果如表7所示。

表7 冻结策略耗时对比			
Table 7 Comparison of freezing strategy time consumption (%)			
冻结策略	YOLOv5 mAP	Faster-RCNN mAP	ResNet50 准确率
Full	78.0	81.22	99.4
LATS	81.1(↑)	82.33(↑)	99.8(↑)

YOLOv5 与 Faster-RCNN 代表检测范式, ResNet50 代表分类范式, 两类任务的评价指标(mAP 与 Acc)存在差异, 因此本文主要关注冻结策略在不同架构下的性能变化趋势而非绝对精度。结果表明, 本文 LATS 方法在不同模型下均能带来性能提升, 说明本文方法对基于卷积的网络结构具有良好的适应性与推广性。

4 结 论

针对小样本任务中训练样本不足导致的特征学习受限问题, 本文分析源域样本与目标域样本的特征分布, 提出了一种基于 CNN-LRP 的层级适应性引导冻结策略生成方法。该方法以 CNN-LRP 的像素级相关性为基础, 引入正向相关比、空间聚焦度和深度惩罚因子, 构建卷积层级的任务适应性评价体系, 并据此生成结构敏感的冻结策略。经过实测实验验证, 本文方法在测试准确率上相比于性能最优的冻结策略提高 3.1%, 参数调优量减少了 34.49%, 并在多种实际模拟场景中表现出更强的泛化能力与鲁棒性, 尤其在强曝光条件下精度提升达 5.8%, 且对于不同卷积网络均具有良好的适应性。该策略通过冻结任务适应性低的卷积层, 引导训练过程聚焦于高适应性卷积层, 减少无效甚至“干扰性”梯度更新, 提升了训练的有效性。本文方法不仅为小样本条件下的冻结策略设计提供了新的解释性支持, 其在任务驱动型层选择上的实用价值也为神经网络结构优化与参数压缩等方向提供了可行的策略设计思路, 具有一定的理论价值与工程实用潜力。需要指出的是, 本文暂未对基于 Transformer 架构的检测模型进行实验。Transformer 架构包含多层自注意力交互与复杂的解码器匹配机制, 直接应用标准 LRP 规则会在 softmax 与注意力矩阵传播过程中产生数值不稳定问题。未来工作将结合此类 Transformer 归因规则, 进一步扩展冻结策略在 Transformer 架构中的可行性验证。

参考文献

- [1] 戴威, 陆小锋, 刘学锋, 等. 基于 Hough 变换和 HSV 彩色空间的电气盘柜状态识别[J]. 计算机测量与控制, 2022, 30(1): 181-187.
DAI W, LU X F, LIU X F, et al. Electrical cabinet status recognition based on Hough transform and HSV color space[J]. Computer Measurement and Control, 2022, 30(1): 181-187.
- [2] 卢正起. 基于机器视觉的船舶航行信号灯图像识别研究[D]. 大连: 大连理工大学, 2023.
LU ZH Q. Research on ship navigation signal light image recognition based on machine vision[D]. Dalian: Dalian University of Technology, 2023.
- [3] LUO D, SHAO J, XU Y, et al. Docking navigation method for UAV autonomous aerial refueling[J]. Science China Information Sciences, 2019, 62: 1-16.
- [4] YANG L, HE Z, ZHAO X, et al. A deep learning method for traffic light status recognition[J]. Journal of Intelligent and Connected Vehicles, 2023, 6(3): 173-182.
- [5] 唐东林, 刘铭璇, 何媛媛, 等. 基于融合度量策略的小样本金属缺陷识别研究[J]. 电子测量技术, 2023, 46(4): 107-113.
TANG D L, LIU M X, HE Y Y, et al. Few-shot metal defect recognition based on fusion metric strategy[J]. Electronic Measurement Technology, 2023, 46(4): 107-113.
- [6] 熊群芳, 林军, 袁希文, 等. 基于深度学习的智轨信号灯检测和识别方法[J]. 机车电传动, 2024(6): 140-148.
XIONG Q F, LIN J, YUAN X W, et al. Detection and recognition method for smart rail signal lights based on deep learning[J]. Electric Drive for Locomotives, 2024(6): 140-148.
- [7] 葛轶洲, 刘恒, 王言, 等. 小样本困境下的深度学习图像识别综述[J]. 软件学报, 2022, 33(1): 193-210.
GE Y ZH, LIU H, WANG Y, et al. A survey of deep learning for image recognition under small sample dilemma[J]. Journal of Software, 2022, 33(1): 193-210.
- [8] 王雷, 何坤, 李宗帅, 等. 基于 BiLSTM-Attention 的迁移学习变工况故障识别方法研究[J]. 电子测量与仪器学报, 2023, 37(7): 205-212.
WANG L, HE K, LI Z SH, et al. Transfer learning method for fault diagnosis under variable working conditions based on BiLSTM-Attention[J]. Journal of Electronic Measurement and Instrumentation, 2023, 37(7): 205-212.
- [9] LEE J, TANG R, LIN J. What would Elsa do? Freezing layers during transformer fine-tuning[J]. ArXiv preprint arXiv:1911.03090, 2019.
- [10] DAR Y, LUZI L, BARANIUK R. Frozen overparameterization: A double descent perspective on transfer learning of deep neural networks[J]. ArXiv preprint arXiv:2209.12755, 2022.
- [11] YOSINSKI J, CLUNE J, BENGIO Y, et al. How transferable are features in deep neural networks? [J].

- ArXiv preprint arXiv:1411.1792, 2014.
- [12] CHEN Z, BADRINARAYANAN V, LEE C, et al. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks[J]. ArXiv preprint arXiv:1711.02257, 2018.
- [13] BROCK A, LIM T, RITCHIE J M, et al. Freezeout: Accelerate training by progressively freezing layers[J]. ArXiv preprint arXiv:1706.04983, 2017.
- [14] SORRENTI A, BELLITTO G, SALANITRI F P, et al. Selective freezing for efficient continual learning[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 3550-3559.
- [15] 李妙彤,王永,肖玲,等. 基于可解释深度学习的电力负荷预测模型[J]. 国外电子测量技术, 2023, 42(4): 66-73.
- LI M T, WANG Y, XIAO L, et al. Power load forecasting model based on explainable deep learning[J]. Foreign Electronic Measurement Technology, 2023, 42(4): 66-73.
- [16] 冯笑,代少升,黄炼. 基于可解释深度学习的单通道脑电跨被试疲劳驾驶检测[J]. 仪器仪表学报, 2023, 44(5): 140-149.
- FENG X, DAI SH SH, HUANG L. Fatigue driving detection across subjects using single-channel EEG based on explainable deep learning[J]. Chinese Journal of Scientific Instrument, 2023, 44(5): 140-149.
- [17] BACH S, BINDER A, MONTAVON G, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation[J]. PLOS ONE, 2015, 10(7): e0130140.
- [18] MONTAVON G, BINDER A, LAPUSCHKIN S, et al. Layer-Wise relevance propagation: An overview[M]. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Cham: Springer, 2019: 193-209.
- [19] ZANG B, DING L, FENG Z, et al. CNN-LRP: Understanding convolutional neural networks performance for target recognition in SAR images[J]. Sensors, 2021, 21(13): 4536.
- [20] YANG S, WU L, WILIEM A, et al. Unsupervised domain adaptive object detection using forward-backward cyclic adaptation[C]. Proceedings of the Asian Conference on Computer Vision (ACCV). Cham: Springer, 2020: 441-456.

作者简介



王振宇, 2023 年于南京航空航天大学获得学士学位, 现为南京航空航天大学硕士研究生, 主要研究方向为图像处理与视觉导航技术。

E-mail: evelyn@nuaa.edu.cn

Wang Zhenyu received her B. Sc. degree from Nanjing University of Aeronautics and Astronautics in 2023. Now she is a M. Sc. candidate at Nanjing University of Aeronautics and Astronautics. Her main research interests include image processing and visual navigation technology.



赵科东 (通信作者), 2015 年于南京航空学院获得学士学位, 2018 年于南京航空航天大学获得硕士学位, 2023 年于南京航空航天大学获得博士学位, 现为南京航空航天大学自动化学院讲师, 主要研究方向为图像/视觉导航、组合导航、航空机载显示技术等。

E-mail: kd_zhao@nuaa.edu.cn

Zhao Kedong (corresponding author) received his B. Sc. degree from Nanjing University of Aeronautics and Astronautics in 2015, M. Sc. degree from Nanjing University of Aeronautics and Astronautics in 2018, and Ph. D. degree from Nanjing University of Aeronautics and Astronautics in 2023, respectively. Now he is a lecturer in the School of Automation, Nanjing University of Aeronautics and Astronautics. His main research interests include image/visual navigation, integrated navigation, aviation airborne display technology and other related fields.



孙永荣, 1991 年于南京航空学院获得学士学位, 1997 年于南京航空航天大学获得硕士学位, 2004 年于南京航空航天大学获得博士学位, 现为正德职业技术学院院长, 南京航空航天大学自动化学院教授, 主要研究方向为视觉导航、多传感器融合智能导航和 GNSS 导航。

E-mail: sunyr@nuaa.edu.cn

Sun Yongrong received his B. Sc. degree from Nanjing Institute of Aeronautics in 1991, M. Sc. degree from Nanjing University of Aeronautics and Astronautics in 1997, and Ph. D. degree from Nanjing University of Aeronautics and Astronautics in 2004, respectively. Now he is dean of Zhengde Polytechnic Institute and Professor in the School of Automation, Nanjing University of Aeronautics and Astronautics. His main research interests include visual navigation, intelligent navigation based on multi-sensor fusion and GNSS navigation.