交通视频监控中异常行为识别算法的轻量化部署策略

王亚勇　马东阳
河南交通投资集团有限公司洛阳分公司　河南　洛阳　471000

摘　要：随着城市化进程的加速和智能交通系统（ITS）的蓬勃发展，对交通视频监控中异常行为（如交通事故、违章停车、行人闯入等）进行实时、精准的自动识别已成为保障道路安全与提升管理效率的关键需求。然而，当前主流的深度学习模型，尤其是基于卷积神经网络（CNN）和Transformer架构的先进算法，普遍存在计算复杂度高、内存占用大、能耗高等问题，难以在资源受限的边缘计算设备（如路侧单元RSU、嵌入式NVR）上实现高效部署。本文聚焦于这一核心矛盾，系统性地探讨了面向交通视频监控场景的异常行为识别算法轻量化部署策略。首先，深入剖析了该任务的独特挑战与约束条件；其次，从模型压缩、知识蒸馏、神经网络架构搜索（NAS）以及软硬件协同优化四个维度，构建了一套完整的轻量化技术体系。研究表明，在显著降低模型参数量与计算量的同时，本方案仍能保持较高的异常行为识别准确率，为智能交通系统的边缘化、实时化部署提供了可行的技术路径。

关键词：交通视频监控；异常行为识别；轻量化；模型压缩；知识蒸馏；边缘计算；智能交通系统

引言

城市交通拥堵、事故频发，给公共安全与社会经济运行带来巨大挑战。传统交通监控依赖人工值守，效率低、成本高，还易因视觉疲劳出现漏检误检。近年来，以深度学习为代表的人工智能技术，为交通视频分析带来革命性突破。大规模神经网络模型可自动从海量视频流中检测车辆、行人，识别追尾、逆行等异常行为，实现事前预警、事中响应与事后追溯的闭环管理。不过，高性能模型多依托强大云端服务器，推理需大量计算资源与带宽。将原始高清视频流全上传云端处理，会产生高通信成本、难以接受的延迟，还有数据隐私泄露风险。所以，把智能分析能力下沉至边缘设备的“边缘智能”成为行业趋势。边缘设备如路侧摄像头、智能信号灯等，算力、内存有限且功耗敏感。在此情形下，如何在保证异常行为识别精度的同时，对复杂深度学习模型“瘦身”“提速”，使其能在边缘设备流畅运行，是亟待解决的核心问题。轻量化部署策略并非单纯牺牲性能换速度，而是要在精度、速度等多维度寻最优平衡，以适应特定场景的严苛约束。

1 相关理论与背景

1.1 交通异常行为识别方法

早期的异常行为识别多基于传统计算机视觉方法，如光流法、背景建模和轨迹聚类。这些方法计算开销相对较小，但鲁棒性差，难以应对复杂多变的交通场景（如光照变化、天气干扰、遮挡等）。随着深度学习的兴起，基于CNN的方法成为主流。Two-stream网络利用RGB帧和光流图分别捕捉空间和时间信息；3D CNN（如I3D, C3D）则直接在时空维度上进行卷积，能更好地建模动态特征。近期，Vision Transformer（ViT）及其变体凭借强大的全局建模能力，在视频理解任务上展现出巨大潜力。然而，这些模型动辄数千万甚至上亿的参数量，远超边缘设备的承载能力。

1.2 模型轻量化技术

模型轻量化是边缘AI领域的研究热点，主要技术路线包括：（1）模型压缩： 通过剪枝、量化和低秩分解等手段，直接减少模型的冗余。剪枝移除不重要的权重或通道；量化将浮点数转换为低比特整数（如INT8），大幅降低内存占用和计算复杂度；低秩分解则用多个小矩阵近似原矩阵。（2）知识蒸馏（KD）： 利用一个庞大而复杂的“教师模型”来指导一个结构简单、参数量少的“学生模型”的学习。学生模型不仅学习真实标签，还学习模仿教师模型的输出（软标签）或中间层特征，从而在较小的体量下获得接近教师模型的性能。（3）轻量级网络设计： 从源头设计高效的网络架构。MobileNet系列引入深度可分离卷积，将标准卷积分解为空间卷积和通道卷积，极大减少了计算量。ShuffleNet则通过通道混洗操作增强特征交互[1]。EfficientNet通过复合缩放统一调整网络的深度、宽度和分辨率，实现了精度与效率的最佳平衡。（4）神经网络架构搜索（ NAS）： 利用自动化算法（如强化学习、进化算法、可微分搜索）在预定义的搜索空间内寻找最优的网络结构。针对特定硬件平台（如移动端、边缘设备）的硬件感知NAS（Hardware-aware NAS）能够直接优化目标设备上的延迟或能耗。尽管上述技术在图像分类等领域取得了显著成

果，但将其直接应用于交通视频异常行为识别仍面临诸多挑战，需要针对性地进行适配与创新。

2 交通视频异常行为识别的轻量化挑战

将通用轻量化技术应用于交通视频监控场景，需克服以下特有挑战：（1）时空依赖性强： 异常行为的本质是时空模式的突变。例如，一次追尾事故涉及前后车辆在短时间内距离的急剧缩短。轻量化模型必须在压缩过程中有效保留关键的时空动态信息，避免因过度简化而导致时序建模能力丧失。（2）小目标与长尾分布： 交通监控画面中，远处的车辆、行人往往是小目标，且异常事件本身属于长尾分布（正常事件远多于异常事件）。轻量化过程可能导致模型对小目标的特征提取能力下降，并加剧类别不平衡问题，使得模型更难识别罕见的异常行为。（3）环境鲁棒性要求高： 交通场景复杂多变，受光照（昼夜、阴晴）、天气（雨、雪、雾）、遮挡（树木、其他车辆）等因素影响严重。轻量化模型通常泛化能力较弱，如何在模型瘦身的同时维持甚至提升其在恶劣环境下的鲁 棒性，是一个巨大挑战[2]。（4）严格的实时性约束： 对于交通事故等紧急事件，识别延迟必须控制在毫秒级，以便及时触发警报或联动措施。这意味着轻量化后的模型不仅要小，还要快，推理速度必须满足视频流的帧率要求（通常 ≥ 15 FPS）。（5）硬件异构性： 边缘设备种类繁多，从低端ARM处理器到配备专用NPU（神经网络处理单元）的高端芯片，其计算架构差异巨大。一种轻量化策略可能在某类硬件上表现优异，但在另一类上效果不佳。

3 轻量化部署策略框架

针对上述挑战，本文提出一个四层递进的轻量化部署策略框架，旨在系统性地解决精度与效率的平衡问题。

3.1 第一层：面向时空建模的模型压缩

传统的模型压缩方法往往将视频帧视为独立的图像进行处理，忽略了其内在的时空关联性。为此，我们采用一种结构化时空剪枝策略。该策略的核心在于评估卷积通道在整个时空维度上的综合重要性，而非孤立地看待单个权重。具体而言，我们通过计算每个通道输出的时空特征图的L1范数或基于梯度的敏感度指标，来衡量其对最终决策的贡献。那些在整个视频片段中持续输出微弱响应的通道被视为冗余，可以被整体移除。这种结构化剪枝不仅大幅减少了模型参数，而且生成的稀疏结构更易于被现代硬件加速器高效执行。在量化方面，我们摒弃了“一刀切”的做法，转而采用渐进式混合量化。对于网络的输入层和输出层，由于它们分别负责原始像素信息的编码和最终语义的解码，对精度极为敏感，我们保留FP32或使用INT16格式；而对于中间的大部分特征提取层，则大胆采用INT8量化。更重要的是，我们引入了时空自适应量化机制，该机制能够根据当前输入视频片段的动态复杂度（例如，通过计算相邻帧间的平均光流强度来衡量）动态调整量化位宽[3]。在车流平稳的简单场景下，系统可以启用更低的位宽以极致压缩；而在发生剧烈冲突的复杂场景下，则自动切换到更高精度模式，确保关键信息不丢失。

3.2 第二层：基于多教师的知识蒸馏

单一的教师模型虽然强大，但其知识可能存在局限性。例如，一个基于3D CNN的模型可能精于捕捉局部的、短时的运动细节，而一个基于Transformer的模型则擅长建模长距离的、全局的时空依赖关系。为了让学生模型获得更全面、更鲁棒的知识，我们设计了一种多教师协同知识蒸馏框架。在此框架中，我们精心挑选了两种不同架构的SOTA模型作为教师：一个是I3D，另一个是Video Swin Transformer。学生模型在训练过程中，不仅要学习真实标签，还要同时模仿这两个教师模型的输出。我们的蒸馏损失函数是一个加权组合，它不仅包含了最终分类logits层的KL散度损失，还融入了中间层特征图的相似性约束。特别地，我们引入了时空注意力蒸馏机制。该机制首先从两个教师模型中分别提取其时空注意力图，这些图揭示了模型在做决策时关注的时空区域。然后，我们引导学生模型去学习这些注意力图的加权平均，从而使其内部的注意力机制能够聚焦于对异常行为判别最关键的时空位置，有效提升了模型在复杂背景下的判别能力。

3.3 第三层：硬件感知的神经网络架构搜索

如果说前两层策略是对现有模型的“改造”，那么这一层则是从源头进行“定制化设计”。我们采用硬件感知的神经网络架构搜索（Hardware-Aware NAS）来自动发现最适合目标边缘设备的网络结构。首先，我们构建了一个灵活且高效的搜索空间，其中包含了多种经过验证的轻量级构建模块，如MobileNetV3中的倒残差块、GhostNet中的Ghost模块以及EfficientNet中的MBConv模块。这些模块的设计初衷就是在最小计算代价下实现最大的特征表达能力。接着，我们将整个异常行为识别任务抽象为一个代理任务—— 动作分类，并在一个小型但具有代表性的子集上快速评估候选架构的性能[4]。最关键的是，我们的奖励函数并非只关注代理任务的准确率，而是将其与在目标硬件平台（例如NVIDIA Jetson Xavier NX）上的实测推理延迟进行联合优化。通过采用可微分的NAS算法（如DARTS），我们能够在连续的架构空间

中高效地进行梯度下降，最终搜索出一个在精度与速度上达到帕累托最优的轻量级骨干网络。

3.4 第四层：软硬件协同优化

再优秀的轻量化模型，若缺乏高效的软件栈支持，也无法在硬件上发挥其全部潜能。因此，软硬件协同优化是轻量化部署的最后一环，也是至关重要的一环。我们选用专为边缘AI优化的推理引擎，如TensorRT或OpenVINO。这些引擎内置了大量针对特定硬件的优化策略。例如，它们能够自动将多个连续的小算子融合成一个大的复合算子，从而减少内核启动开销和内存读写次数；它们还能智能地复用中间激活值的内存，极大降低峰值内存占用。此外，我们对整个视频处理流水线进行了精心设计。视频解码、图像预处理（如缩放、归一化）、模型推理和后处理（如非极大值抑制、事件判定）等步骤被组织成一个多阶段的流水线，并通过多线程或异步I/O技术并行执行。这样一来，当一个线程在进行模型推理时，另一个线程可以同时处理下一帧的解码和预处理，有效隐藏了I/O等待时间，显著提升了系统的整体吞吐量，确保了端到端的实时性。

4 轻量化部署与系统集成挑战

尽管前述轻量化策略在理论上为交通视频监控中的异常行为识别提供了可行路径，但其工程落地仍面临多重系统性挑战。首先，交通场景动态演化易引发“概念漂移”，需构建高效数据闭环，通过边缘端筛选困难样本、上传元数据与关键片段，在带宽与隐私约束下实现模型的持续迭代。其次，路侧设备高度异构（如海思、瑞芯微、Jetson等芯片平台），难以统一部署，需依赖容器化与ONNX等标准接口，并针对主流硬件维护多版本模型库以实现动态适配。第三，系统必须具备强鲁棒性与容错能力，支持网络中断下的本地自治运行、健康状态监控及自愈机制，并引入不确定性估计以规避因输入异常导致的误判。最后，安全与隐私合规至关重要，须贯彻“隐私设计”原则，通过硬件安全模块、视频实时脱敏、联邦学习等手段，保障数据全生命周期安全，防范模型窃取与对抗攻击，确保系统可信可控。

5 结语

本文针对交通视频监控中异常行为识别算法在边缘设备部署难题，提出系统性轻量化策略，通过结构化压缩、知识蒸馏、架构搜索及软硬件协同优化，在模型大小、推理速度和识别精度上取得卓越平衡。未来研究将沿三方向深入：一是探索模型在边缘端的在线持续学习能力，让其能增量吸收新知识，避免频繁回传云端重训练；二是融合雷达等传感器信息，构建多模态轻量级融合模型，提升系统在极端天气下的鲁棒性；三是探索联邦学习框架，在保护边缘节点数据隐私前提下，协同训练出更强大、泛化能力更强的全局轻量级模型。轻量化是连接前沿AI算法与现实应用的桥梁，随着技术演进，更安全、高效、智能的城市交通未来可期。

参考文献

[1]杜楠.智能交通视频监控系统的分析与应用研究[J].人民公交,2025,(14):40-42.

[2]徐天珍,智慧交通视频监控管理云平台V1.4.17.安徽省,安徽逸路安科技股份有限公司,2022-04-08.

[3]冯睿,邓袁钦.道路交通视频监控设备数据结构化效能评估及组合应用[J].湖南交通科技,2020,46(04):135-138.

[4]田竞辰.高速公路视频监控技术研究[J].中国新技术新产品,2024,(19):79-81.