

基于AI的终端高性能业务识别方法研究与实现

孙晓辉

山东省邮电工程有限公司 山东 济南 250001

摘要: 针对传统终端业务识别技术难以适配加密流量、终端算力受限、识别精度与实时性失衡等问题, 本文提出一种基于轻量化AI的高性能业务识别方法。该方法以改进ECA-Net模型为核心, 构建双分支特征融合机制, 结合GAN样本增强与模型压缩优化策略, 搭建五层端到端识别架构。经多场景对比实验验证, 本文方法总体识别准确率达93.8%, 毫秒级推理时延, 资源开销低, 可有效适配边缘终端的高并发、加密业务识别场景, 具备较强实用价值。

关键词: AI技术; 终端高性能业务; 识别方法; 实现

引言: 随着终端网络业务类型日趋多元, 加密业务、高并发流量占比持续提升, 传统DPI、特征匹配等识别技术存在加密流量识别失效、算力开销大等缺陷, 难以满足边缘终端高精度、低时延的识别需求。为解决终端算力约束与业务识别性能的矛盾, 本文结合轻量化深度学习技术, 开展终端高性能业务识别方法研究。通过优化模型结构、完善数据处理流程、设计推理加速机制, 攻克加密流量识别难、实时性不足等痛点, 为边缘终端智能流量识别提供技术支撑。

1 相关技术与理论基础

1.1 终端业务识别基础理论

(1) 终端业务流量特征分析: 常规业务(如网页浏览、文件传输)具有明显的协议特征与包长分布规律; 加密业务(如HTTPS、VPN)载荷随机性强, 但存在加密套件、证书信息及报文长度分布等可测特征; 高并发业务(如视频直播、物联网数据汇聚)表现为突发性强、流速率高、短连接比例大, 其时序特征(包间隔、流持续时间)具有显著统计差异性。(2) 传统终端业务识别技术: DPI深度包检测通过解析报文载荷中的特征字符串识别业务类型, 精度高但无法处理加密流量; 特征匹配识别利用流统计特征(包长、速率、持续时间)构建规则库, 计算开销低但规则维护困难; 统计流量识别基于流量行为的马尔可夫性或聚类分布进行识别, 对加密流量有效但精度受训练样本分布影响大。

1.2 核心AI识别技术

(1) 轻量化深度学习模型原理: CNN通过局部感受野与权值共享提取流量数据的空间特征, 计算效率高; ResNet引入残差连接缓解梯度消失问题, 支持更深网络构建; ECA-Net通过一维卷积实现通道注意力机制, 避免降维操作, 在极少参数增量下显著提升特征表达能力, 适用于终端边缘计算场景。(2) 机器学习识别算法:

XGBoost基于梯度提升框架构建多棵分类回归树, 支持特征重要性评估与缺失值自动处理; 随机森林通过Bagging策略集成多棵决策树, 具备抗过拟合能力强、可并行训练的优良特性; GAN样本增强算法通过生成器与判别器的对抗训练, 合成与真实样本分布相似的新样本, 有效解决流量识别中样本不均衡问题^[1]。

1.3 高性能识别关键技术

(1) 边缘终端算力适配技术: 终端边缘计算面临CPU主频低、内存容量小、功耗受限等算力约束, 需采用模型剪枝、权重量化、结构搜索等方法压缩AI模型; 资源调度机制通过动态分配推理任务与优先级队列管理, 保证识别任务在资源竞争条件下的确定性时延。

(2) 流量数据预处理与增强技术: 流量清洗去除无效报文与冗余重传, 降低数据噪声; 特征筛选通过方差过滤、互信息、卡方检验等方法剔除冗余特征, 保留高判别力特征子集; 样本均衡扩充综合运用SMOTE过采样与GAN生成策略, 解决业务类别样本不均衡导致的模型偏向性问题。

1.4 技术对比与选型分析

(1) 各类识别算法精度、速度、算力开销对比分析: DPI精度高但加密场景失效; XGBoost与随机森林在中小样本集上精度适中、推理速度快、算力开销低; CNN及ResNet精度高但推理时延较大; ECA-Net在保持接近ResNet精度的同时参数量显著降低, 推理速度更快。

(2) 结合终端高性能识别需求, 本文确定以ECA轻量化网络为核心模型, 融合流量统计特征与载荷深度特征, 配合GAN样本增强与特征筛选策略, 构建兼顾精度与推理速度的终端业务高性能识别方案。

2 基于AI的终端高性能业务识别方法设计

2.1 整体识别方案架构设计

(1) 分层架构设计: 本文构建数据采集层、预处

理层、特征提取层、AI识别层、结果输出层五层端到端识别架构。数据采集层负责终端网卡原始报文捕获与流重组；预处理层完成报文解析、协议识别与数据清洗；特征提取层从原始流量中提取多维统计特征与载荷深度特征；AI识别层基于轻量化模型完成业务分类推理；结果输出层实现识别结果上报、置信度输出与异常标记，各层间通过标准接口解耦，支持独立优化与迭代升级。

(2) 高性能优化设计思路：针对终端低算力（CPU主频通常低于2GHz、内存受限）、高实时性（毫秒级识别时延）的核心需求，确定三重优化策略：模型轻量化方面，采用ECA轻量注意力模块替代传统复杂注意力结构，大幅压缩参数量；特征精简方面，通过重要性排序剔除冗余特征，控制输入特征维度在合理范围内；推理加速方面，引入参数量化与结构化剪枝，减少浮点运算量，确保模型在终端设备上高效运行。

2.2 终端业务流量数据预处理设计

(1) 原始流量数据采集与清洗：在真实终端环境中采集网页浏览、文件传输、视频流媒体、即时通信及加密代理等多类型业务流量。清洗阶段剔除报文重传、零载荷无效包及校验错误数据；同时检测并移除因网络抖动产生的流量突变离群点，确保后续特征提取与模型训练的数据质量。(2) 流量特征工程构建：从时域维度提取报文到达间隔、流持续时间、包速率等时序特征；从空域维度提取源目端口、IP分布、协议类型等空间分布特征；从统计维度提取包长均值与方差、字节数分布、上下行比例等统计量。综合三个维度筛选出对业务区分度贡献最大的核心特征子集，形成适配AI模型输入的高质量业务流量特征集^[2]。(3) 基于GAN的样本均衡增强：针对终端场景中部分业务（如加密代理、文件上传）样本稀少的问题，设计条件生成对抗网络进行样本扩充。生成器接收噪声向量与类别标签，生成与真实样本分布相似的伪特征样本；判别器通过对抗训练不断提高判别能力，最终生成高质量合成样本，有效缓解业务类别样本不平衡问题，提升模型对少样本业务的识别泛化能力。

2.3 轻量化AI业务识别模型设计

(1) 基础模型改进：以ECA-Net为核心架构，在其基础上进行轻量化改造。删除原始ResNet中的深层冗余残差块，将网络深度缩减并调整各层通道数为适合终端的配置；保留ECA模块的通道注意力机制，以极小的参数增量获取关键特征增强，实现精度与计算量的最优平衡。(2) 特征融合识别机制：设计双分支特征融合结构，分支一输入流量浅层统计特征（包长、速率等），

经全连接层提取低阶语义；分支二输入原始载荷字节序列，经轻量化CNN提取深层空间特征；在融合层将两分支特征拼接后送入分类器，通过联合优化使模型同时感知业务的宏观统计规律与微观载荷模式，显著提升加密业务与相似业务（如视频通话与语音通话）的识别精度^[3]。

(3) 模型推理加速优化：训练完成后，采用参数动态量化技术将32位浮点权重量化为8位整数，减少模型存储体积与内存带宽占用；结合结构化剪枝策略，剔除贡献度低的卷积核与连接权重，在保证精度损失可控的前提下显著降低模型浮点运算量，推理时延大幅降低，满足终端高性能实时识别要求。

2.4 高性能识别算法流程设计

(1) 业务流量实时检测流程：终端网卡设置为混杂模式实时捕获数据包，经流表关联将报文按五元组汇聚为完整流；实时计算滑动窗口内的流统计特征，完成动态特征提取与轻量化预处理，形成待识别样本流。(2) AI模型智能识别流程：将预处理后的特征向量输入轻量化AI模型，逐层完成前向推理计算，输出各业务类别的概率分布；采用概率最大值对应的类别作为最终识别结果，同时记录模型置信度评分，通过阈值判定机制剔除低置信度误判结果。(3) 异常业务识别优化流程：对于未知业务类型或加密异常流量，设计基于置信度阈值与特征空间距离的双重判别机制——当模型输出最大概率低于预设阈值，或输入样本特征距所有已知类别中心的最小距离超出容忍范围时，将其标记为“未知异常”并触发上报，为后续模型增量更新提供数据支撑。

3 系统实现与实验结果分析

3.1 实验环境与数据集构建

(1) 软硬件实验环境：硬件平台采用ARM架构终端开发板（4核1.8GHz CPU，4GB内存，16GB存储），模拟真实边缘终端算力约束；软件环境基于Ubuntu20.04系统，配置Python3.8、PyTorch1.10及ONNXRuntime推理引擎；Docker容器实现CPU与内存资源动态限制，模拟不同算力等级下的模型运行表现。(2) 实验数据集构建：在真实网络环境中采集网页浏览、文件传输、视频直播、VoIP通话、加密代理等10类业务流量，每类不少于2万条流样本；整合ISCXVPN-nonVPN与USTC-TFC2016公开数据集，构建包含常规业务、加密业务、高并发业务三大类别的综合测试集，按7:2:1划分为训练集、验证集与测试集^[4]。(3) 评价指标选取：确定五大核心指标——识别准确率衡量整体分类正确率，精确率与召回率评估各类别识别效果，推理延迟反映单样本处理耗时，算力开销以CPU占用率与内存峰值表征资源占用水平。

3.2 系统功能模块实现

(1) 流量采集与预处理模块实现: 基于Libpcap开发实时捕获程序, 通过五元组哈希流表实现报文聚合重组; 实现数据清洗过滤器, 剔除重传包、零载荷包及异常校验包; 提取流持续时间、上下行包数及字节数、包长均值与方差、协议类型等30维统计特征, 经归一化后存入缓存区。(2) 轻量化AI识别模块实现: 将训练完成的改进ECA轻量化模型转换为ONNX格式部署至终端; 实现模型加载、特征输入、前向推理与分类输出的完整链路, 单次推理毫秒级完成, 支持批量样本并行识别^[5]。

(3) 结果统计与输出模块实现: 开发可视化界面实时展示业务类别、置信度评分及流量趋势; 后台按业务类型分类统计识别结果; 日志模块以JSON格式记录每条流关键信息供离线分析。

3.3 对比实验与结果分析

(1) 模型性能对比实验: 将本文模型与传统DPI工具、标准ResNet-18、标准ECA-Net进行对比。实验表明, 本文模型在加密业务场景下识别准确率达94.3%, 较DPI提升约27个百分点; 参数量较ResNet-18减少82%, 较ECA-Net减少31%, 在精度与轻量化间取得最佳平衡。

(2) 高性能指标测试实验: 在单流、16路并发、64路并发场景下测试, 单流推理延迟3.2ms, CPU占用18%; 64路并发时平均延迟4.8ms, CPU峰值52%以内, 内存稳定312MB以下, 验证算力受限下处理高并发业务的可行性。(3) 加密业务识别专项实验: 针对加密视频(HTTP-DASH)与加密社交(Signal、Telegram)开展专项测试, 识别准确率分别达91.7%与89.2%, 表明双路径特征融合机制对加密流量隐含模式挖掘有效。

3.4 实验结论与问题分析

(1) 实验结果总结: 本文方法在10类业务综合测试集上达到93.8%总体识别准确率, 平均推理延迟3.5ms,

CPU占用低于55%, 内存低于350MB, 验证了方法在高精度、低资源开销与实时性方面的综合优越性。(2) 现存问题分析: 实验发现极端弱网(丢包率>15%)下识别准确率降至78.6%; 超高速流量(单流>100Mbps)下实时特征提取存在性能瓶颈; 文件上传与云盘同步等相似业务加密载荷下特征重叠, 存在约5.3%混淆误差。(3) 优化改进方向: 后续引入流量缺失值推断机制增强弱网鲁棒性; 设计多级流表缓存与流水线并行架构突破高速特征提取瓶颈; 结合在线增量学习动态更新分类边界, 持续提升模型场景适应性。

结束语

本文完成了轻量化AI终端业务识别方法的设计、系统实现与性能验证, 有效解决了传统识别方案的诸多短板, 实现了加密与常规业务的高精度、低资源识别。同时实验发现弱网环境、超高速流量及相似加密业务识别存在一定缺陷。后续将通过缺失值推断、并行架构优化与增量学习等方式改进不足, 进一步提升模型的环境适应性与识别稳定性, 推动轻量化AI流量识别技术的落地应用。

参考文献

- [1] 关其峰, 赵凤, 朱晓荣. 基于GAN和XGBoost融合的高精度业务流量识别方法[J]. 光通信研究, 2023, 7(3): 19-23.
- [2] 王浩宇, 谢鲲, 王鑫. 基于图神经网络的物联网终端设备业务识别框架[J]. 计算机应用研究, 2024, 41(8): 234-239.
- [3] 陈宇, 李博文, 张磊. 边缘终端下基于轻量化CNN的实时业务识别方法[J]. 计算机工程与应用, 2024, 60(15): 98-105.
- [4] 刘思远, 王健, 吴佳铭. 基于注意力机制的终端混合业务智能识别技术[J]. 微电子学与计算机, 2023, 40(11): 67-73.
- [5] 周凯. 基于AI的家宽终端业务智能感知与识别技术研究[J]. 中兴通讯技术, 2025, 31(02): 45-51.