

基于指令调优与 LLM 的非独立同分布 加密流量分类方法

陈 光, 李炳龙, 于世龙

(信息工程大学, 河南 郑州 450001)

摘要: 针对实际网络环境中流量数据呈现出显著的非独立同分布(Non-IID)特性, 提出一种融合大语言模型(LLM)与指令调优的分类方法。该方法借助 LLM 对结构化数据的理解能力来实现加密流量分类, 通过将流量数据转换为图结构以显式表征流间逻辑关联, 并在此基础上构建两阶段训练框架。首先, 图结构指令调优使 LLM 掌握流量图语义理解能力; 其次, 流量任务指令调优将 LLM 的推理能力引导至分类任务。二者形成能力递进的协同关系。在 USTC-TFC2016 与 CIC-IDS-2017 数据集上的实验显示, 该方法宏平均 $F1$ 分数分别达到 0.945 和 0.968, 较主流基线模型提升显著。结果证明, 该方法能够有效学习流量图中的复杂依赖关系, 其分层学习机制为处理非独立同分布流量数据提供可行路径。

关键词: 加密流量分类; 非独立同分布; 大语言模型; 指令调优; 图结构

中图分类号: TP393.08

文献标识码: A

文章编号: 1671-0673(2026)01-0097-08

Non-Independent and Identically Distributed Encrypted Traffic Classification Method Based on Instruction Tuning and LLM

CHEN Guang, LI Binglong, YU Shilong

(Information Engineering University, Zhengzhou 450001, China)

Abstract: To address the problem that the traffic data in the actual network environment shows significant non-independent and identical distribution (Non-IID) characteristics, a classification method that integrates large language models (LLM) with instruction tuning is proposed. In this method, the understanding capability of LLM for structured data is used to achieve encrypted traffic classification, and traffic data is converted into graph structures to explicitly represent logical relationships among flows. Based on this, a two-stage training framework is constructed. Firstly, the LLM is enabled to comprehend the semantics of traffic graphs by using graph structures instruction tuning. Subsequently, the reasoning capability of LLM is guided to the classification task through the traffic task instruction tuning. These two stages form a collaborative relationship with progressive capability enhancement. Experiments on USTC-TFC2016 and CIC-IDS-2017 datasets show that macro-average $F1$ scores reach 0.945 and 0.968 by the using proposed method, respectively, demonstrating significant improvement over those of mainstream baseline models. The results indicate that complex dependencies in traffic graphs can be effectively learned by using the proposed method, and a feasible approach for handling Non-IID traffic data is provided through the hierarchical learning mechanism.

Key words: encrypted traffic classification; Non-IID; LLM; instruction tuning; graph structures

网络加密流量的分类是网络管理与安全分析中的一项基础性任务。随着 TLS、QUIC 等加密协议的广泛部署,传统基于载荷深度检测的技术路径已难以适用,研究重心逐渐转向基于机器学习的分类方法^[1]。现有技术路线主要围绕两类方法展开:1)基于手工设计的流统计特征结合传统分类器^[2];2)利用深度学习模型从数据包序列中自动学习特征表示^[3]。近年来,亦有研究尝试通过图神经网络(Graph Neural Networks, GNN)对流量间的关联关系进行建模^[4]。

通过对现有文献的梳理发现,当前加密流量分类方法普遍基于数据独立同分布的前提假设^[5]。然而在实际网络环境中,流量数据呈现出显著的非独立同分布特性:一方面,流量内部存在时序依赖关系,不同流量间具有复杂的逻辑关联;另一方面,流量特征分布会随时间、空间和应用演变而发生动态变化^[6]。这种特性导致现有方法在真实场景中面临模型泛化能力不足的挑战。

针对上述问题,本研究旨在探索一种能够有效处理非独立同分布特性的加密流量分类方法。具体研究内容包括:提出一种将网络流量转化为图结构的数据表示方法,以显式刻画流内时序关系和流间逻辑关联;构建基于大语言模型(Large Language Models, LLM)的指令调优框架,通过分阶段的调优策略使模型逐步掌握流量图结构理解能力和分类任务执行能力;设计系统的实验方案,验证所提方法在典型应用场景下的有效性。

本研究的主要价值体现在3个方面:在方法论层面,将 LLM 与指令调优引入加密流量分析领域,为处理复杂的网络数据依赖关系提供了新的技术路径;在技术实现层面,提出的流量图表示方法能够更完整地保留网络交互的上下文信息;在应用层面,设计的层次化指令调优框架为适应动态变化的网络环境提供了可行的解决方案。

1 相关理论与技术基础

1.1 加密网络流量概述

加密网络流量指采用 TLS/SSL 等加密协议对传输内容进行保护的数据流。其广泛部署使得基于深度包检测的传统分类方法因载荷不可读而失效,研究重点转向基于机器学习的分类范式^[7-8]。主要分为如下两类。

1)基于统计特征的方法:提取包长、时序、流持续时间等元数据特征,使用决策树、随机森林等

模型进行分类。该方法不依赖载荷内容,但难以捕获复杂模式。

2)基于深度学习的方法:采用卷积神经网络(Convolutional Neural Network, CNN)、循环神经网络(Recurrent Neural Network, RNN)、长短期记忆网络(Long Short-Term Memory, LSTM)等网络从原始包序列中自动学习特征。虽然效果显著,但多数方法基于独立同分布假设,且可解释性差,难以有效处理实际网络中的复杂依赖关系。

1.2 非独立同分布特性

网络流量本质上是非独立同分布的^[9],主要体现在如下两类。此特性使得在静态数据集上训练的模型在实际部署时泛化性能显著下降。如何建模上下文依赖与适应分布变化成为核心挑战。

1)样本依赖性:流量单元间存在强关联。流内数据包遵循严格协议时序(如 TCP 握手);流间存在逻辑因果,如用户访问行为触发的 DNS、TLS、HTTPS 流构成 1 个 Burst(在网络中,极短的时间窗口内,突然出现的一系列在逻辑上紧密相关的数据包或网络流,这正是非独立同分布特性的典型体现),形成完整上下文。

2)分布不一致性:流量特征分布随时空、用户群体及新型应用出现而动态变化,导致模型面临分布偏移问题。

1.3 LLM 与指令调优

LLM 是基于 Transformer 架构的海量参数模型,通过预训练获得通用的语言理解和生成能力。指令调优是指在 LLM 基础上,使用包含任务指令和对应输出的数据集进行有监督微调的关键技术,目的是让其学会在接收到自然语言指令时,能够理解指令并完成任务。其核心机制是通过将不同任务统一转化为“指令—响应”的文本生成模式,使模型学会遵循指令格式并准确执行特定任务^[10]。

在加密流量分类中,LLM 与指令调优构成能力适配的协同关系。基础 LLM 提供通用的序列建模和模式识别能力,而指令调优作为领域适配的关键环节,分阶段将这种通用能力转化为领域专用技能。具体而言,首先通过图结构指令调优使 LLM 掌握流量图的语义理解能力,其次通过流量任务指令调优将这种理解能力引导至分类决策任务。这种分层调优策略有效解决 LLM 与专业领域之间的语义鸿沟,实现从通用语言模型到流量分析专家的能力转化,为处理非独立同分布数据提供新的技术路径。

1.4 相关领域研究进展

近年来,随着LLM和GNN技术的快速发展,加密流量分析领域涌现出诸多创新性研究成果。

在LLM与指令调优方面,研究者开始探索将预训练语言模型应用于加密流量分析的新范式。其中,指令调优技术通过将分类任务转化为“指令—响应”的文本生成模式,有效提升了模型在非独立同分布场景下的适应能力^[11]。研究表明,基于自监督指令调优的框架能够在同分布和异分布数据集上均取得显著的性能提升,表现出优越的泛化能力。

在图结构学习领域,最新研究突破传统GNN的局限。流注意力机制通过将基尔霍夫第一定律融入图注意力计算,为网络流量等具有守恒特性的图数据建模提供了新思路^[12]。同时,图流匹配模型通过反应—扩散方程整合邻域信息,在保持生成质量的同时提升了训练效率。这些进展为构建更精确的流量图表示奠定了理论基础。

在跨模态对齐技术方面,结构感知的软提示调优方法通过类型感知嵌入和跨模态对齐模块,有效融合了图结构语义与文本特征。这种方法以线性计算成本实现图与文本的深度交互,为LLM理解复杂的流量图结构提供了可行的技术路径^[13]。

这些研究成果从不同角度推动了加密流量分析技术的发展,为本研究的框架设计和方法创新提供了坚实的理论支撑和实践指导。

2 方法设计

本文提出一种基于LLM与指令调优的创新方法,命名为“ITCEN-LLM”(Instruction-Tuning Classification Encrypted Network LLM)。该方法的核心思想是通过将网络流量抽象为图结构以显式捕获其依赖关系,并设计分层次的指令调优算法将通用LLM逐步转化为领域专用分类器。具体实现包含3个核心组件:首先,设计流量图结构构建方法,通过定义时序边与逻辑边将原始流量转换为结构化图表示;其次,提出两阶段图结构指令调优算法,分别通过图结构指令调优和流量任务指令调优实现模型能力的递进式提升;最后,部分流量分类任务调优实现,将图序列化描述与任务指令相结合,实现端到端的分类决策。该方法通过算法设计与系统实现的结合,为破解非独立同分布场景下的加密流量分类难题提供了完整的技术解决方案。总体框架如图1所示。实线单箭头表示主数据流;虚线单箭头表示任务数据/指令流;实线双箭头表示交互式训练流程。

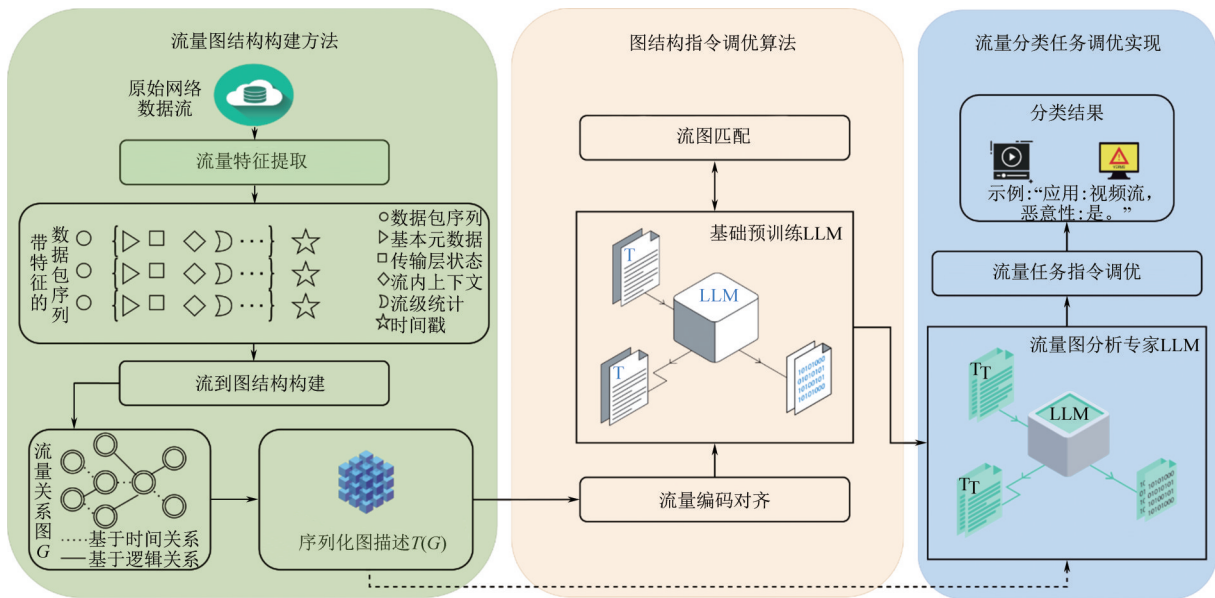


图1 ITCEN-LLM方法总体框架

2.1 流量图结构构建方法

通过定义节点映射规则和边构建算法,将原始网络流量数据转化为结构化的图表示。

具体实现分为两个阶段:首先,通过特征提取器从数据包中抽取多维度特征并完成特征融合;

其次,基于时间关联和逻辑关系构建连接边,形成包含完整拓扑信息的流量关系图。

2.1.1 流量特征提取

流量特征提取负责从原始网络流中捕获并融合多维度、多粒度的特征信息。

1) 输入与预处理。

输入: 原始网络流量数据, 通常以 pcap 文件的形式输入。

流会话重组: 首先, 根据五元组 (源 IP、目的 IP、源端口、目的端口、传输层协议) 对数据包进行会话重组, 得到独立的网络流 F_i 。

TCP 报文段重组: 对于 TCP 流, 基于序列号与确认号对 TCP 分段进行重组, 还原出完整的应用层消息单元。此步骤对于理解应用层语义至关重要。

2) 多维度特征提取。

数据包级特征: 对每个重组后的数据包 P_j , 提取其关键属性, 构成特征向量 f_j 。

timestamp: 数据包到达的相对或绝对时间戳。

direction: 传输方向 (上行: 0; 下行: 1)。

payload_length: 应用层载荷长度。

protocol_flags: 从 TCP/IP 头中提取的标志位集合 (如 SYN、ACK、FIN、PSH)。

ttl: 生存时间。

流级统计特征: 对整个网络流 F_i 计算统计特征, 如总数据包数、平均包大小、流量持续时间、字节熵等, 这些特征可以作为该流中所有数据包的全局上下文。

3) 特征融合与输出。

将数据包级特征与流级统计特征进行拼接或加权融合, 为每个数据包 P_j 生成 1 个最终的综合特征向量 $\mathbf{x}_j$ 。输出: 对于每个网络流 F_i , 输出 1 个带特征的数据包序列 $S_i = [(\mathbf{x}_1, t_1), (\mathbf{x}_2, t_2), \dots, (\mathbf{x}_n, t_n)]$, 其中 t 为时间戳, 该序列保留了流内的时序信息。

2.1.2 流到图结构构建

流到图结构构建旨在将多个相关的流会话整合到一个统一的图结构中, 以建模流间的相关性。

1) 图定义。构建 1 个有向图 $G=(V, E)$ 。其中: V 是节点集合; E 是边集合。

2) 节点生成。将上个步骤中输出的每 1 个数据包 P_j 映射为图 G 中的 1 个节点 v_j , 节点 v_j 的特征向量即为 $\mathbf{x}_j$ 。

3) 边构建 (关系定义)。边的构建是捕获非独立同分布特性的核心, 本文定义时序边和逻辑边两类主要关系。

时序边 E_{temp} : 在同 1 个流 F_i 内部, 按照数据包的到达顺序, 在相邻的数据包节点 v_j 和 v_{j+1} 之间建立有向边, 边类型标记为 [NEXT_IN_FLOW]。此边用于建模流内的时序依赖性。

逻辑边 E_{logic} : 在不同流之间, 基于网络协议语

义和主机关联日志, 建立跨流的逻辑连接边。这是捕获 Burst 流间相关性的关键。

请求—响应边: 例如, 在 1 个 HTTP 请求流与对应的 HTTP 响应流中, 将请求数据包节点与响应数据包节点连接, 边类型标记为 [RESPONDS_TO]。

因果依赖边: 例如, 将 DNS 查询流中的请求节点与后续 TCP 连接流中的 SYN 包节点连接, 边类型标记为 [TRIGGERED_BY]。

会话关联边: 对于 TLS/SSL 等多报文握手协议, 将属于同一握手阶段的数据包节点连接, 边类型标记为 [HANDSHAKE_WITH]。

4) 图序列化。为了将图 G 输入 LLM, 需将其线性化为 1 个令牌序列, 采用结构化的文本描述法, 模板如图 2 所示。输出: 图 G 的文本化描述 $T(G)$, 作为后续 LLM 处理的输入。

```
[GRAPH_START]
<NODE_1> [PACKET] direction: {dir}, length: {len}, flags:
{flags}, ..., -> [EDGE_TYPE] <NODE_2>
<NODE_2> [PACKET] direction: {dir}, length: {len}, ..., ->
[EDGE_TYPE] <NODE_3>
...
[GRAPH_END]
```

图 2 结构化的文本描述法

2.2 图结构指令调优算法

本节旨在解决 LLM 与图结构数据之间的语义鸿沟问题。通过设计层次化训练策略, 首先采用掩码语言建模实现基础语义对齐, 其次通过流图匹配任务提升结构推理能力, 最后使 LLM 具备理解流量图语义的能力, 为下游任务奠定基础。

2.2.1 流量编码对齐

此步骤是一个自监督的预训练过程, 目标是将序列化图描述 $T(G)$ 的词汇与结构模式嵌入到 LLM 的语义空间中。

1) 任务设计: 采用掩码语言建模任务。

2) 训练数据构建: 从大规模未标注的流量图中自动生成训练样本。对于每个序列 $T(G)$, 随机掩码掉 15% 的令牌。这些令牌包括如下: 关键属性值: 如 direction:[MASK]、length:[MASK]; 关系类型: 如 →[MASK]<NODE_X>; 节点标识符: 如 <NODE_[MASK]>。

3) 训练目标: 最小化 LLM 预测被掩码令牌的交叉熵损失。通过此过程, LLM 学习到 [RESPONDS_TO]、uplink、length 等令牌在流量图上下文中的具体语义, 实现了文本空间与流量图特征空间的表

示对齐。

2.2.2 流图匹配

在编码对齐的基础上,通过设计特定的指令任务,让LLM进行更深层次、面向图结构的推理。

1)任务设计与数据构建:构建1个指令数据集 $D_{\text{graph}} = \{(I_i, G_i, A_i)\}_{i=1}^N$, 其中: I_i 是指令; G_i 是图; A_i 是期望的答案; N 为总样本数。任务类型包括如下。

图属性查询:指令如“图中上行数据包的数量是多少?”,训练LLM进行图遍历与统计。

关系推理:指令如“找出所有响应节点<NODE_5>的节点”,训练LLM理解边的关系语义。

子图识别与描述:指令如“描述以节点<NODE_1>为中心,两跳内的子图结构”,训练LLM进行局部结构理解与自然语言概括。

2)训练过程:使用标准的下一个令牌预测目标,在数据集 D_{graph} 上对已完成编码对齐的LLM进行监督微调。损失函数为

$$L_{\text{loss_graph}} = - \sum_{t=1}^{|A|} \ln P(a_t | I, G, a < t) \quad (1)$$

式中: a_t 为答案序列 A 中的第 t 个令牌; $P(\cdot)$ 为条件概率。

3)输出:经过本阶段训练,LLM 成为一个“流量图分析专家”,能够准确理解并执行对流量图结构的各种查询和推理指令。

2.3 流量分类任务调优实现

本节聚焦于将图结构专家模型适配到具体分类任务。通过构建任务指令集并融入少样本提示与思维链机制,引导模型将已习得的图结构知识转化为分类决策能力,实现从“理解图结构”到“执行分类任务”的能力跃迁。

1)任务特定指令数据集构建。构建最终的指令数据集 $D_{\text{task}} = \{(I_i, G_i, L_i)\}_{i=1}^M$, 其中: L_i 为任务相关的标签(如应用类型、恶意性); M 为总样本数。指令 I_i 需明确任务目标。一是应用识别:基于所提供的流量图,将其分类到视频流、网页浏览、文件传输、语音通信等类别之一,仅输出类别名;二是恶意流量检测:分析此流量图,判断其是否表现为恶意软件通信,回答“是”或“否”。

2)高级提示技术集成。少样本提示:在指令 I_i 中,动态添加 k 个($k=1,2,3$)已标注的($G_{\text{example}}, L_{\text{example}}$)样本作为上下文示例,以提供任务范式和增强模型在非独立同分布数据上的泛化能力。思维链提示机制:对于复杂任务,通过修改指令格式,要求

模型在输出最终分类结果前,先生成基于流量图特征的推理过程。例如,指令调整为:“请先分析该流量的数据包大小分布和交互时序特征,然后给出最终分类。”相应的期望输出为:“该流量呈现持续下行大包和周期性上行小包模式,符合视频流特征。结论:视频流。”

3)训练与优化。以经过第2.2节调优的LLM作为初始化模型。在数据集 D_{task} 上进行监督微调,优化目标为最大化给定指令和图上下文下输出正确答案的概率。损失函数为

$$L_{\text{loss_task}} = - \sum_{l=1}^{|L|} \ln P(l_l | I, G, l < l) \quad (2)$$

式中, L 为模型需要生成的目标文本序列,其具体内容由任务设计决定,可以是简单的类别名,也可以是包含思维链的复杂文本。通过最小化基于 L 计算的损失 $L_{\text{loss_task}}$,模型被训练成能够根据指令和图上下文,自动生成符合期望格式和内容的答案,从而完成分类任务。

3 实验设计与结果分析

3.1 实验环境与数据集

3.1.1 实验环境

硬件环境。GPU: NVIDIA® RTX™ 4090 24 GB GDDR6X×1; CPU: AMD Ryzen™ 9 7950X(16核心32线程); 内存: 128 GB DDR5 4 800 MHz; 存储: 2 TB PCIe® 4.0 NVMe™ SSD。

软件环境。操作系统: Ubuntu 22.04 LTS; 深度学习框架: PyTorch 2.1.0, Transformers 4.35.2; 图处理库: NetworkX 3.1、PyTorch Geometric 2.4.0; 编程语言: Python 3.10。

RTX 4090 的 24 GB 显存足以满足 LLaMA-7B 模型全参数微调(BF16精度)的需求,其强大的算力能保证高效的训练速度。128 GB 内存确保大型流量数据集图构建过程的流畅性,该配置在成本与性能间取得了优异平衡。

3.1.2 数据集

选用公开的加密流量数据集 USTC-TFC2016 与 CIC-IDS-2017,并对其进行重构以适配非独立同分布场景下的分类任务。

USTC-TFC2016:该数据集包含 20 种不同应用的流量,其中:10 种为恶意软件流量;10 种为正常应用流量。选取其中流量交互模式丰富的 8 类(如 BitTorrent、Weibo、Minecraft、Zeus 等)进行实验。

CIC-IDS-2017:该数据集包含常见的网络攻击

流量与正常流量。选取“Botnet”、“DDoS”、“Web Attack”与“Benign”4类流量,以验证框架在恶意流量检测上的泛化能力。

为模拟真实的非独立同分布环境,未对数据集进行随机打散,而是依据时间戳,将前70%的流量会话作为训练集,后续15%作为验证集,最后15%作为测试集。此举保证了训练集与测试集在时间分布、网络状态及背景流量上存在固有差异,更符合现实世界场景。最终,构建的数据集统计信息如表1所示。

表1 实验数据集统计信息

数据集	类别数	训练流 会话数	验证流 会话数	测试流 会话数	备注
USTC-TFC2016	8	28 500	6 100	6 100	应用识别
CIC-IDS-2017	4	185 000	39 600	39 600	恶意 流量检测

3.1.3 基线模型与参数设置

为全面评估本框架的性能,与以下4类代表性基线模型进行对比。

1)机器学习(Machine Learning, ML)方法。C4.5算法:决策树模型,使用从流中提取的统计特征(如包长、间隔的均值、方差等)进行训练。

2)深度学习(Deep Learning, DL)方法。1D-CNN:直接处理数据包大小序列。LSTM:用于建模流量序列的长期依赖关系。深度森林(Deep Forest, DF):一种基于堆栈式自编码器的深度学习方法,使用流统计特征作为输入。

3)GNN。图卷积网络(Graph Convolutional Network, GCN):在本文构建的流量关系图上进行训练,以验证图结构本身的有效性。

4)LLM。Zero-Shot:使用未经调优的基础预训练LLM(LLaMA-7B),通过零样本提示直接进行分类。Fine-Tuned:在序列化的流量文本数据上直接对LLM进行全参数微调,但不经过本框架的图结构指令调优阶段。

参数设置:本框架中,基础LLM选用LLaMA-7B。受经济配置的单卡环境所限,采用梯度累积策略、混合精度训练、激活检查点来稳定训练。图结构指令调优阶段学习率设为 10^{-5} ,批量大小设置为4,梯度累积步数为8。流量任务指令调优阶段学习率设为 5×10^{-6} ,批量大小为8,梯度累积步数为4。均采用AdamW优化器。此设置在RTX4090的24 GB显存限制下,实现了最佳的训练效率与稳定性。

3.2 对比实验与评价指标

3.2.1 对比实验设计

设计3组对比实验,从不同维度验证框架的优越性。

1)总体性能对比:将本框架ITCEN-LLM与第3.1.3节中列出的所有基线模型在两个数据集上进行全面对比,评估其综合分类性能。

2)消融实验分析:为验证框架中各个核心组件的贡献,设置以下消融模型。

ITCEN-LLM w/o Graph:移除图结构,仅将数据包序列及其特征作为文本直接输入LLM进行指令调优。

ITCEN-LLM w/o G-Instruction:移除图结构指令调优阶段,直接在基础LLM上进行流量任务指令调优。

ITCEN-LLM w/o T-Instruction:保留图结构指令调优,但最终阶段改为使用线性分类头进行分类,而非指令调优。

3)鲁棒性分析:通过改变训练集与测试集的时间划分比例(如8:2、6:4),观察各模型性能随分布偏移程度加大的衰减情况,评估其鲁棒性。

3.2.2 评价指标

采用机器学习中通用的准确率、精确率、召回率、 $F1$ 分数、宏平均 $F1$ 分数等分类评价指标,以全面衡量模型性能:

$$R_{\text{accuracy}} = \frac{N_{\text{TP}} + N_{\text{TN}}}{N_{\text{TP}} + N_{\text{TN}} + N_{\text{FP}} + N_{\text{FN}}} \quad (3)$$

$$R_{\text{precision}} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FP}}} \quad (4)$$

$$R_{\text{recall}} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FN}}} \quad (5)$$

$$F1 = \frac{2 \times R_{\text{precision}} \times R_{\text{recall}}}{R_{\text{precision}} + R_{\text{recall}}} \quad (6)$$

$$F1_{\text{macro}} = \frac{1}{C} \sum_{i=1}^C F1_i \quad (7)$$

式中: N_{TP} 表示真正例; N_{TN} 表示真负例; N_{FP} 表示假正例; N_{FN} 表示假负例; C 为总类别数; i 为类别索引。

3.3 实验结果分析

3.3.1 总体性能对比分析

采用宏平均 $F1$ 分数作为核心评价指标,能够有效避免因类别样本数量不均衡而导致的评估偏差,确保模型在所有类别上的性能都得到公正的衡量。本文方法与其他方法的性能对比如表2所示。本文方法在两个数据集上均取得了最高的宏平均 $F1$ 分数,证明了其在经济配置下依然能充分发挥效能,原因分析如下。

表 2 性能对比

方法	模型	USTC-TFC2016(应用识别)					CIC-IDS-2017(恶意流量检测)				
		准确率	精确率	召回率	F1	F1 _{macro}	准确率	精确率	召回率	F1	F1 _{macro}
ML	C4.5	0.829	0.826	0.821	0.823	0.821	0.894	0.889	0.885	0.887	0.887
	1D-CNN	0.861	0.857	0.852	0.854	0.854	0.919	0.914	0.910	0.912	0.912
	LSTM	0.875	0.869	0.865	0.867	0.868	0.932	0.927	0.923	0.925	0.925
DL	DF	0.886	0.881	0.877	0.879	0.879	0.938	0.933	0.929	0.931	0.931
	GCN	0.898	0.892	0.889	0.890	0.891	0.947	0.942	0.938	0.940	0.940
LLM	Zero-Shot	0.470	0.452	0.448	0.450	0.463	0.525	0.508	0.515	0.511	0.518
	Fine-Tuned	0.909	0.904	0.899	0.901	0.902	0.952	0.947	0.943	0.945	0.945
本文方法		ITCEN-LLM	0.952	0.948	0.943	0.945	0.975	0.970	0.966	0.968	0.968

1)图结构与 LLM 的协同效应。GCN 的良好表现验证了图结构的有效性。ITCEN-LLM 的进一步提升,证明 LLM 通过指令调优,能够比 GCN 更深度地挖掘和利用图中的复杂语义与逻辑关系。

2)分阶段指令调优的优势。直接微调的 Fine-Tuned 性能已不错,但本框架通过图结构指令调优这一预备阶段,让模型先“学会看图”,再“学习解题”,取得了更优的效果。

3.3.2 消融实验分析

基于 USTC-TFC2016 数据集,消融实验结果如表 3 所示。1)图结构是关键基石。移除图结构(w/o Graph)导致性能大幅下降 4.9 个百分点,这表明将流量构建为图是有效建模非独立同分布特性的核心。2)图结构指令调优是成功前提。跳过此阶段(w/o G-Instruction)的性能损失达 6.3 个百分点,这表明没有经过专门的“识图”训练,LLM 难以理解自定义的图序列化格式,流量编码对齐与流图匹配是必不可少的。3)指令调优优于传统分类头。使用线性分类头(w/o T-Instruction)的性能低于完整框架 2.7 个百分点,说明指令调优能更好地激发和利用 LLM 固有的推理能力。

表 3 消融实验

模型	准确率	精确率	召回率	F1	F1 _{macro}
ITCEN-LLM (完整框架)	0.952	0.948	0.943	0.945	0.945
ITCEN-LLM w/o Graph	0.902	0.898	0.892	0.894	0.896
ITCEN-LLM w/o G-Instruction	0.888	0.884	0.878	0.880	0.882
ITCEN-LLM w/o T-Instruction	0.925	0.920	0.916	0.918	0.918

3.3.3 鲁棒性分析

基于 USTC-TFC2016 数据集,通过调整训练集与测试集的划分比例(从 7:3 调整为 5:5)来模拟分

布偏移,不同模型的鲁棒性分析如表 4 所示。ML 性能下降 10.1 个百分点,DL 下降 7.6 个百分点,而本文方法仅下降 4.8 个百分点。这种优势源于图结构对本质交互模式的捕捉能力,以及指令调优中少样本提示的快速适应机制,使其在非独立同分布场景下具有更强鲁棒性。进一步分析表明,框架在处理未知应用类型时也表现出良好的泛化能力。在仅提供少量样本的情况下,模型能够通过指令中的上下文学习快速适应新的流量模式。

表 4 不同模型在分布偏移下的鲁棒性分析

方法	模型	F1 _{macro} (7:3)	F1 _{macro} (5:5)
ML	C4.5	0.821	0.720
DL	DF	0.879	0.803
本文方法	ITCEN-LLM	0.945	0.897

3.3.4 效率分析

在经济配置下,记录典型的训练时间。完成 ITCEN-LLM 在 USTC-TFC2016 数据集上的全部训练(包括图结构指令调优和流量任务指令调优)约需 38 h。作为对比,GCN 模型的训练仅需约 2 h。尽管训练耗时更长,但 ITCEN-LLM 带来的显著性能提升证明了其投入是值得的,且该时间在通用科研设备可接受的范围内。

3.3.5 与先进方法的对比分析

1)任务设定与数据划分的不同。基于上述数据集,目前虽然有不少研究表现出更好的分类效果,但多是基于独立同分布的数据划分假设,即在训练前对全部数据进行随机打散^[14-15]。虽然能有效评估模型的学习能力,但这种设定忽略了真实网络流量在时间上的分布漂移和上下文依赖性,简化了分类任务。本研究的核心目标是评估模型在非独立同分布场景下的泛化能力,因此采用时间感知的划分方法(按时间戳将前 70% 的流量会话作为训练集,后续 15% 作为验证集,最后 15% 作

为测试集)。尽管这一严苛设定导致所有模型的绝对性能下降,但它更能真实反映模型在实际网络环境中的适用性。

2)模型复杂性与效率的权衡。当前,确实有在其特定的任务设定下取得了优异分类效果的方法,但同时也带来了更高的计算开销^[1]。本研究提出的框架旨在探索一种基于通用LLM的新范式,是在性能、泛化能力和模型通用性之间寻求平衡。

3)研究焦点的差异。本文方法并非追求在特定数据集上的最高指标,而是验证“图结构表示”与“LLM指令调优”相结合这一技术路径在解决加密流量非独立同分布问题上的有效性。实验结果表明,本框架在该挑战性设定下显著优于一系列代表性基线模型,为后续研究提供了新的思路。

4 结束语

针对非独立同分布网络加密流量分类的关键问题,提出一种基于LLM与指令调优的创新解决方法。提出一种分层图结构构建方法,将原始网络流量转化为包含时序关系和逻辑关联的图表示;设计一种两阶段指令调优机制,通过图结构指令调优和流量任务指令调优的递进式训练,实现LLM从通用语义理解到领域专业能力的转化;构建完整的加密流量分类框架,在多个数据集上验证方法的有效性。然而,本研究仍存在局限性。首先,计算开销高昂,难以满足实时在线部署的低延迟要求;其次,实验所使用的数据集中包含的部分应用与特定地区的实际网络环境存在差异,部分应用在中国的网络环境中并非主流或受到访问限制,这在一定程度上影响了模型在实际场景中的泛化能力评估;最后,图构建过程仍依赖人工定义的关联规则,未能实现完全的自动化。

未来工作聚焦方向:1)研究模型压缩与加速技术,如量化与知识蒸馏,以推动框架的实际应用落地;2)构建更具代表性的本土化数据集,并探索跨域与少样本学习机制,以增强模型对未知流量和真实复杂环境的适应能力;3)探索基于LLM的自动化图结构生成方法,减少对人工先验知识的依赖。

参考文献:

- [1] 于治平,刘彩霞,刘树新,等.基于机器学习的网络流量分类综述[J].信息工程大学学报,2023,24(4):447-453;483.
- [2] DARWISH R, ROY K. Comparative analysis of federated

learning, deep learning, and traditional machine learning techniques for IoT malware detection[C]//Proceedings of the 2025 IEEE 4th International Conference on AI in Cybersecurity. Piscataway, USA: IEEE, 2025. DOI: 10.1109/ICAIC63015.2025.10849203.

- [3] MELVIN A A R, KATHRINE J W, JEYABOSE A, et al. A deep learning model leveraging time-series system call data to detect malware attacks in virtual machines[J]. International Journal of Computational Intelligence Systems, 2025, 18(1): No.58.
- [4] 张国明.基于图卷积神经网络的大规模软件定义网络流量预测模型[J].微电子学与计算机,2024,41(4): 96-103.
- [5] 张帅华,张淑芬,周明川,等.基于半监督联邦学习的恶意流量检测模型[J].计算机应用,2024,44(11):3487-3494.
- [6] GORSHENIN A, KOZLOVSKAYA A, GORBUNOV S, et al. Mobile network traffic analysis based on probability-informed machine learning approach[J]. Computer Networks, 2024, 247: No.110433.
- [7] 唐博麟,王晨飞,江帆,等.基于网络流时空序列的加密流量分类[J].计算机应用与软件,2024,41(3):297-302.
- [8] 陈子涵,程光,徐子恒,等.互联网加密流量检测、分类与识别研究综述[J].计算机学报,2023,46(5):1060-1085.
- [9] 顾永跟,高凌轩,吴小红,等.非独立同分布下联邦半监督学习的数据分享研究[J].计算机工程,2024,50(6): 188-196.
- [10] 黄君豪,朱锦文,向宗元,等.大语言模型引导的文本摘要技术与系统[J].计算机应用,2024,44(增刊1): 29-33.
- [11] LIN X J, XIONG G, GOU G P, et al. Respond to change with constancy: instruction-tuning with LLM for non-I.I.D. network traffic classification[J]. IEEE Transactions on Information Forensics and Security, 2025, 20: 5758-5773.
- [12] PLETTENBERG P, KHÖLER D, SICK B, et al. Flow-attentional graph neural networks[DB/OL]. (2025-09-27) [2025-10-05]. <https://arxiv.org/abs/2506.06127>.
- [13] FENG R J, PEARCE H, LIGUORI P, et al. CGP-tuning: structure-aware soft prompt tuning for code vulnerability detection[J]. IEEE Transactions on Software Engineering, 2025, 51(9): 2533-2548.
- [14] ZHANG X X, WANG Q, QIN M Y, et al. Enhanced few-shot malware traffic classification via integrating knowledge transfer with neural architecture search[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 5245-5256.
- [15] LIU Z, LI T, YANG Y, et al. Encrypted malicious traffic detection based on spatial-temporal graph learning[C]//Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. New York, USA: ACM, 2023: 345-357.

(编辑:李志豪)