ET-BERT: A Contextualized Datagram Representation with Pre-training Transformers for Encrypted Traffic Classification

WWW 2022

Xinjie Lin, Gang Xiong, Gaopeng Gou, Zhen Li, Junzheng Shi, **Jing Yu***

2022.03.24

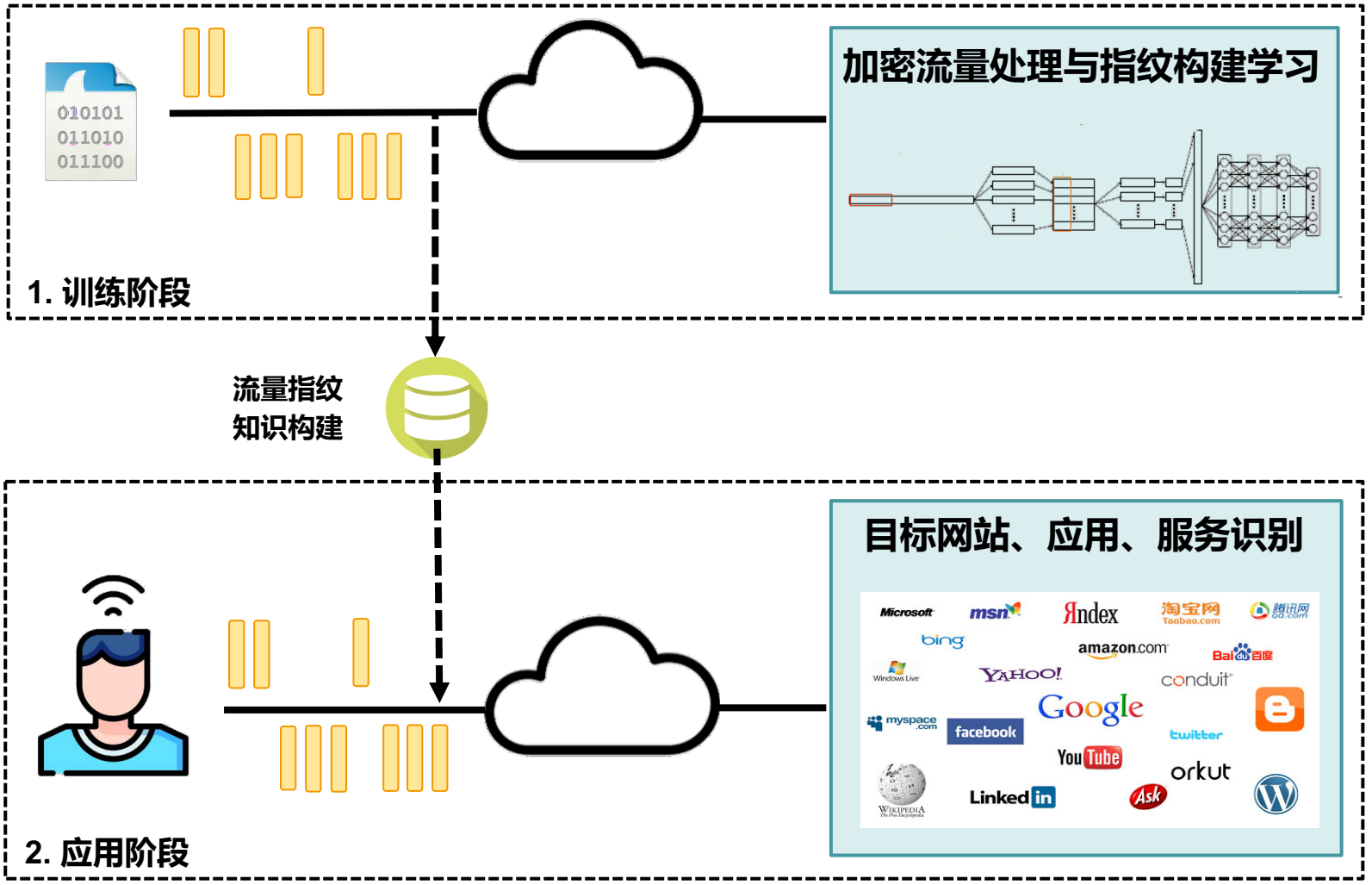
中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

中国科学院大学
University of Chinese Academy of Sciences

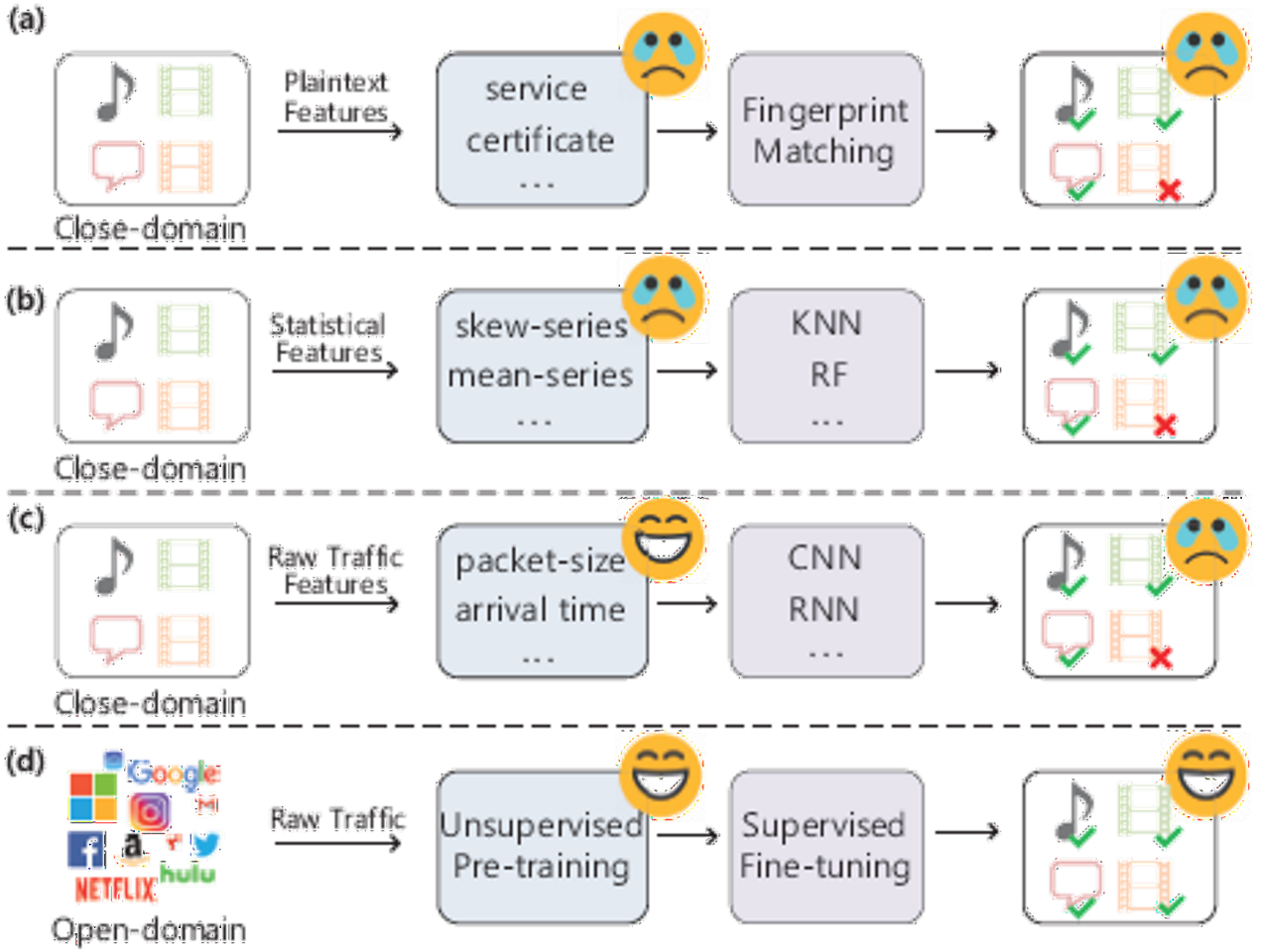
目录

- 研究背景
- 模型介绍
- 实验分析
- 总结展望

加密网络流量分类


010101
011010
011100
加密流量处理与指纹构建学习
1. 训练阶段
流量指纹
知识构建
目标网站、应用、服务识别
2. 应用阶段

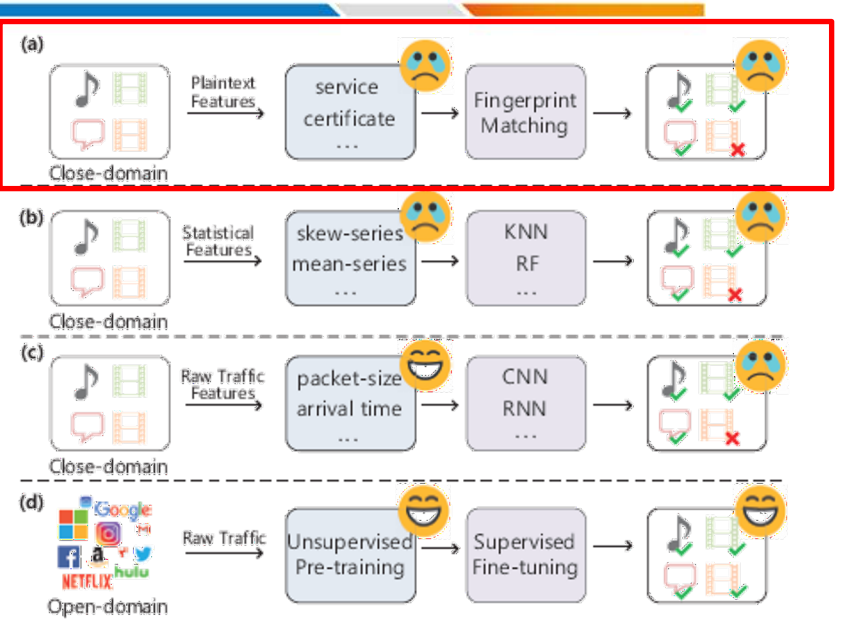
加密网络流量分类


(a)
Plaintext Features
service
certificate
...
Fingerprint Matching
Close-domain
(b)
Statistical Features
skew-series
mean-series
...
KNN
RF
...
Close-domain
(c)
Raw Traffic Features
packet-size
arrival time
...
CNN
RNN
...
Close-domain
(d)
Raw Traffic
Unsupervised Pre-training
Supervised Fine-tuning
Open-domain

相关工作

基于规则的流量指纹构建

- 利用加密流量的**字段组合、排序**或者**固定模式**等作为指纹进行模式匹配。
  - 字段组合特征
    - [Kim et al., APNOMS 2015]: 构建Certificate、session ID和IP对应关系列表。
    - [Shbair et al., ICDCSW 2016]: 对比SNI和IP对应的域名信息。
  - 排列顺序特征
    - [Husák et al., EURASIP JIS 2016]: 采用ciphersuite list和HTTP中的user-agent 。
  - 固定模式特征
    - [Papadogiannaki et al.,RAID2018]:正则匹配固定模式，如包出现频率或包所在的位置。



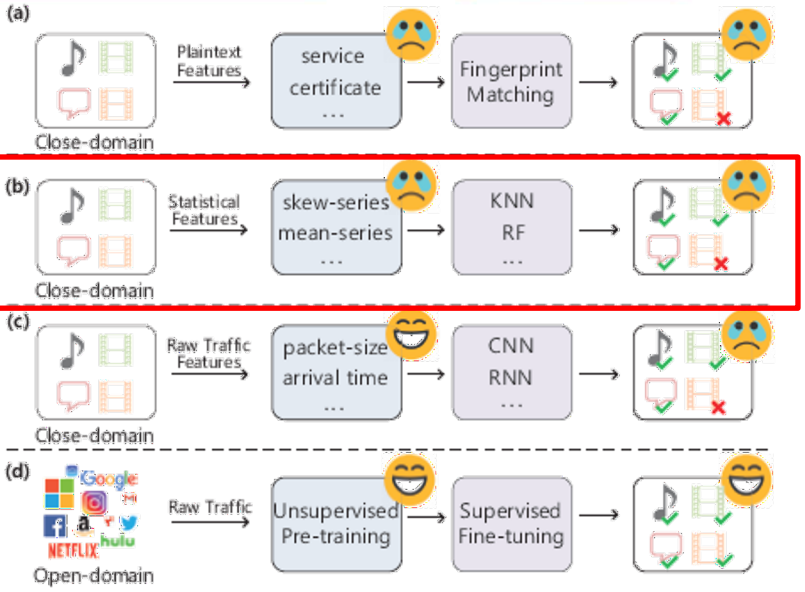
- **优点**：轻量级的识别加密流量。
- **缺点:**

1. 需要**人工分析**海量级流量，选择具有区分性的字段特征或组合。
2. 仅可以对**已提取的规则**进行匹配识别。
3. 容易被人工拼接或恶意伪造字段的流量**绕过**，导致**高误报率**。

相关工作

基于人工特征的流量指纹构建

- 通过挖掘流量的**不同维度信息**特性结合**统计方法**，构建有效指纹特征。
  - 基于**数据包/流**的特征
    - **代表工作 1**
      - [Jamie et al., USENIX Security 2016]：C2S和S2C方向的**数据包数量总和**、**均值**和**标准差**，数据包**方向**序列等多类特征构建指纹，对Tor场景下的网站进行识别。
    - **代表工作 2**
      - [Shen et al., IWQoS 2016]：使用数据包**证书消息类型及长度**作为特征，通过二阶马尔可夫矩阵建模转移概率，对12类主流应用进行识别。
    - **代表工作 3**
      - [Anderson et al., KDD 2017]：使用数据包**包长和时间**的**最大最小均值方差**、**等分包长度块**构建基础特征，同时引入client hello的**握手元信息**包括cipher suite等进行指纹构建，对恶意流量进行识别。
    - **代表工作 4**
      - [Sengupta et al., WWW 2019]：通过将**数据包**报文的**比特序列**进行傅里叶变换，生成频谱向量，结合密码套件和**包大小**统计特征，对移动应用进行识别。



- **优点**：流量指纹特征构建易于理解。
- **缺点**：

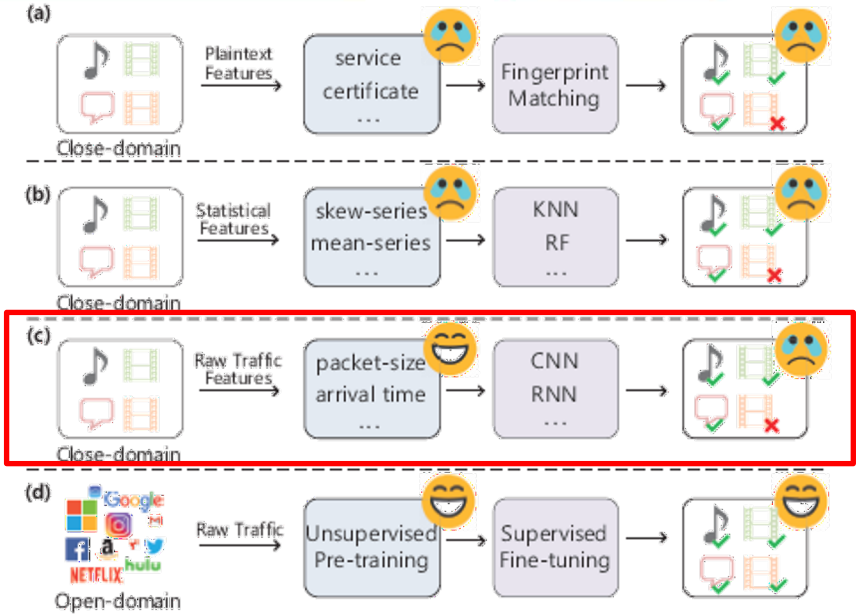
1. 特征指纹需要**人工设计**，需要依赖专家经验和专业知识。

2. 特征指纹难以适应多场景且保持高性能的迁移，缺乏**普适性、泛化性**。

相关工作

基于原始流量特征的指纹构建

- 通过深度学习模型自动学习加密流量原始数据并构建指纹。
  - **基于原始序列的表示学习**
    - **代表工作 1**
      - [Sirinam et al., CCS 2018]：通过**卷积神经网络**对**流方向序列**进行特征自动提取、生成和分类，实现对网站加密流量的特征表示。
    - **代表工作 2**
      - [Payap et al., CCS 2019]：通过**三重网络学习**流量**包大小序列**的嵌入表征，以高度区分不同类别的样本，并且支持小规模样本的训练。
    - **代表工作 3**
      - [Liu et al., INFOCOM 2019]：通过**双向门循环单元**机制联合重构损失和分类损失对**包大小序列**进行表示学习，该不依赖流的长度规模且支持其他序列信息的特征表示。
    - **代表工作 4**
      - [Shen et al., TIFS 2021]：提出TIG对流的交互关系进行图建模并基于**图表示学习**表征会话**包大小序列**的关联，相比传统神经网络模型和特征构建方法具有更好的流量识别能力。
    - **代表工作 5**
      - [Wu et al., ICC 2021]：提出**双向长短期记忆网络**和**注意力机制**表征带方向的**包大小流序列**，在包含标准Web信息的加密流量中具有更好的应用识别能力。



- **优点**：特征构建不依赖人工选择。
- **缺点**：

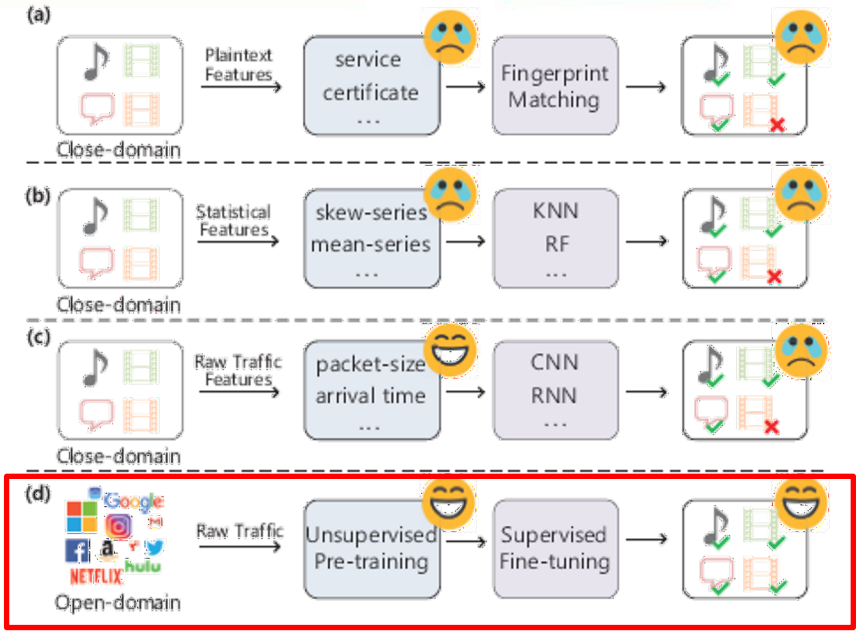
1. 依赖**大规模**的**标注**流量数据，模型性能受限于数据规模。

2. 流量输入的**信息选择单一化**难以应用到多场景任务。

相关工作

基于原始流量数据的指纹表征

- 通过深度学习模型自动提取加密流量原始数据的通用表示。
  - **基于原始序列的表示学习**
    - **代表工作 1**
      - [Li et al., IWQoS 2018]：使用切分原始流量报文的分段作为单元数据，联合循环神经网络和注意力机制对每个分段报文的重要特征进行表示。
    - **代表工作 2**
      - [He et al., ITU 2020]：使用预训练网络作为无标注流量的自监督监督学习，对加密流量数据包报文进行表征并迁移到3个不同场景下进行流分类。
    - **代表工作 3**
      - [Lotfollahi et al., Soft Computing 2020]：使用流量原始报文的字节编码向量，结合一维卷积网络和长短期记忆网络，实现高效表示流量且捕捉到远距离的报文间关系。
    - **代表工作 4**
      - [Lin et al., Computer Networks 2021]：使用流量原始报文的字节编码向量，结合一维卷积网络和长短期记忆网络，实现高效表示流量且捕捉到远距离的报文间关系。



- **优点**：自动化学习流量表征，不依赖人工与特征选择。
- **缺点:**

1.未充分利用**无标注**流量数据，如何构造合适的无标注流量特征学习任务是难点。

2. 依赖**大规模标注**流量数据。

相关工作

规则指纹构建

- IP地址，协议端口号
- 握手连接信息（SNI、证书字段等）

人工特征指纹

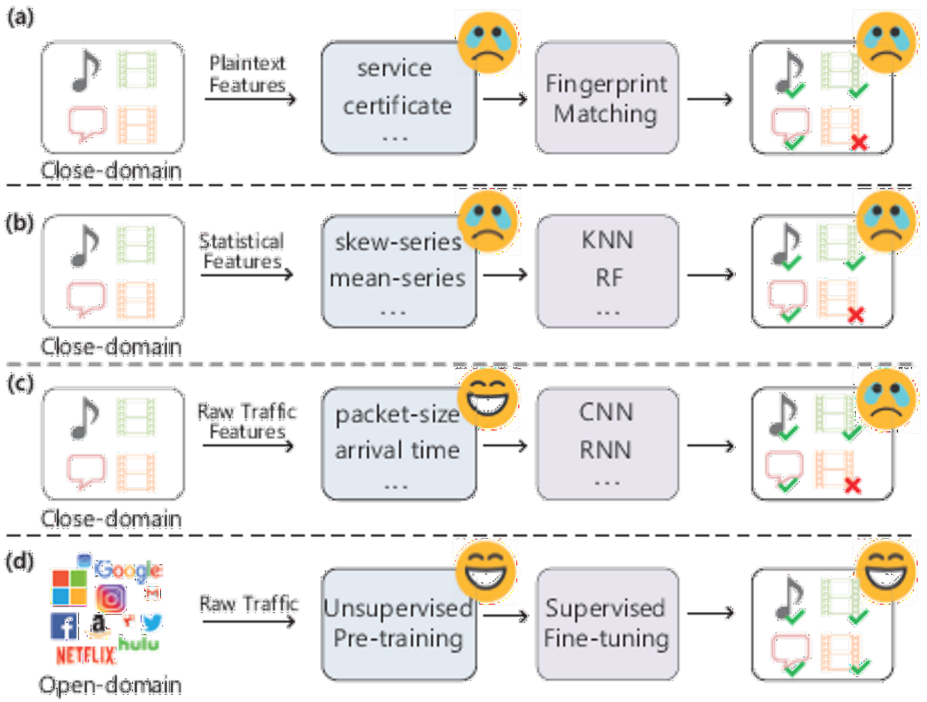
- 包长、时间戳、方向最大、最小、均值、方差等统计特征

原始流量特征

- 包长、时间戳、方向等原始序列特征
- 基于深度学习模型生成流量表征

原始流量数据

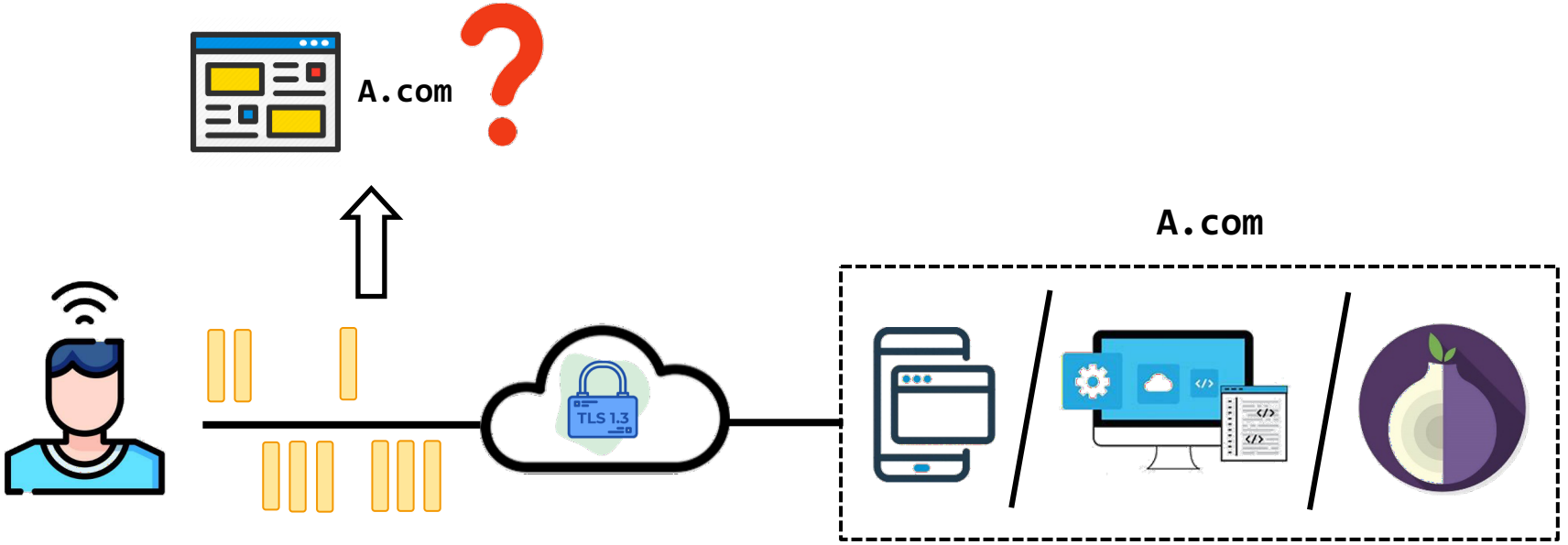
- 数据报文等原始流量信息
- 基于无监督学习模型生成流量通用表征并有监督学习泛化到不同场景



Note: 加密流量的明文信息逐渐减少，流量规模庞大且标注困难，传统研究工作逐渐被淘汰

背景：传统流量分类研究存在的挑战

- 流量明文信息被加密（ECH, ESNI, …）
- 匿名网络、混淆隧道的特征隐匿（MTU, …）
- 开放环境（私有加密协议等未知加密服务与应用的存在）
- 标注流量的技术成本和社会成本增加（标注依据缺失，隐私保护法律）
- …



研究动机

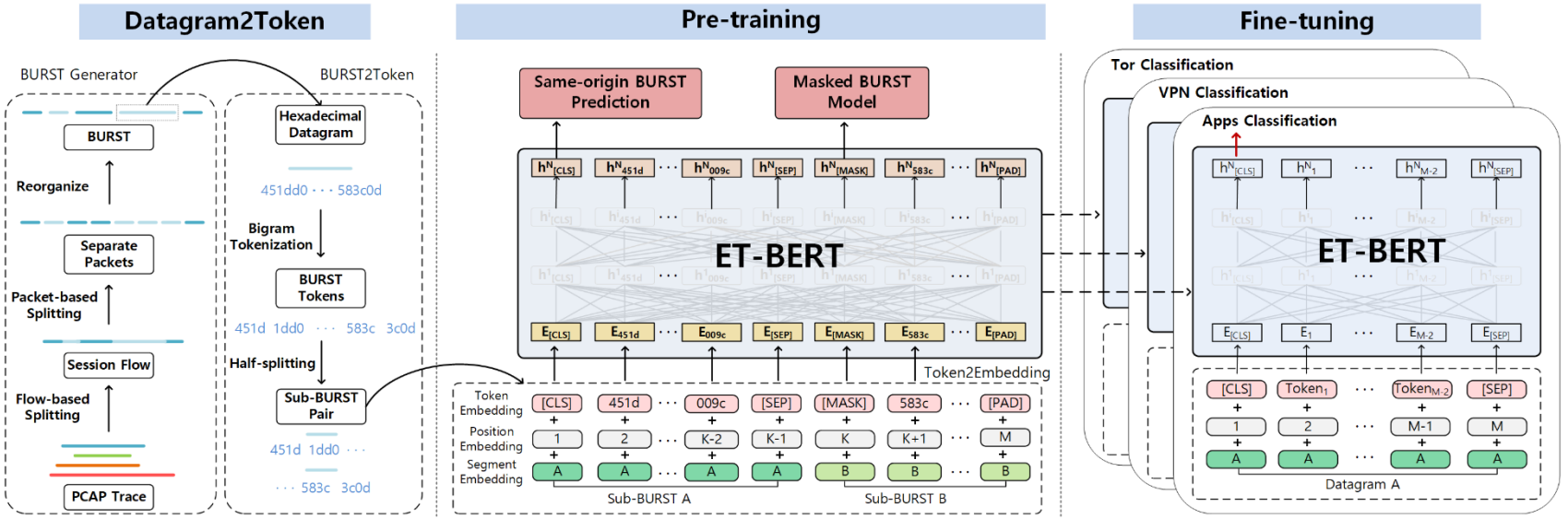
研究问题

- 特征构建**繁杂**（依赖专家经验 / 大规模标注流量数据）
- 多场景迁移的**泛化**能力弱，迁移代价高

研究目标

- 构建一个基于无标注流量数据的全加密流量表征学习模型
  - 不依赖人为经验构建特征工程
  - 不依赖大规模的标注流量数据
- 实现在多种场景下流量识别的迁移学习
  - 考虑加密流量的通用表征，不考虑具体的协议、应用等
  - 使用少量标注数据微调，快速迁移到不同场景

研究框架



ET-BERT：Encrypted Traffic Bidirectional Encoder Representations from Transformer

- **Datagram2Token** — 加密流量报文预处理与符号编码，作为预训练的输入数据
- **Pre-training** — 自监督预训练阶段，学习加密流量内容与结构的上下文信息
- **Fine-tuning** — 微调训练阶段，根据具体场景的流量调整预训练好的通用表征

研究内容：流量报文转化为符号向量

原始流量 Trace

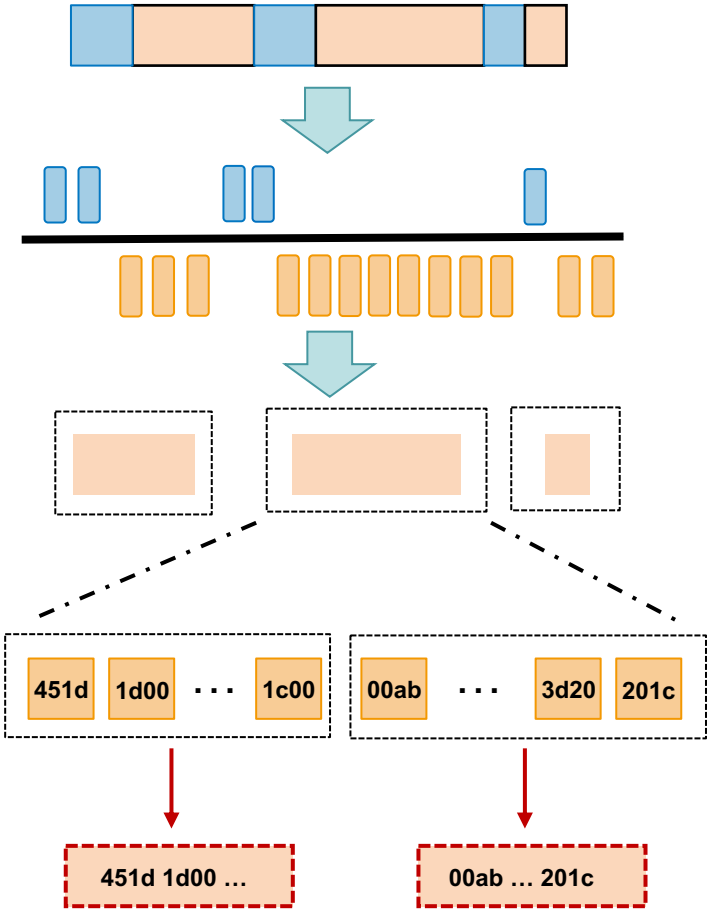
双向会话流的数据包

连续单向数据包组成BURST

Bigram 编码BURST

Byte Pair Encoder

切分 BURST



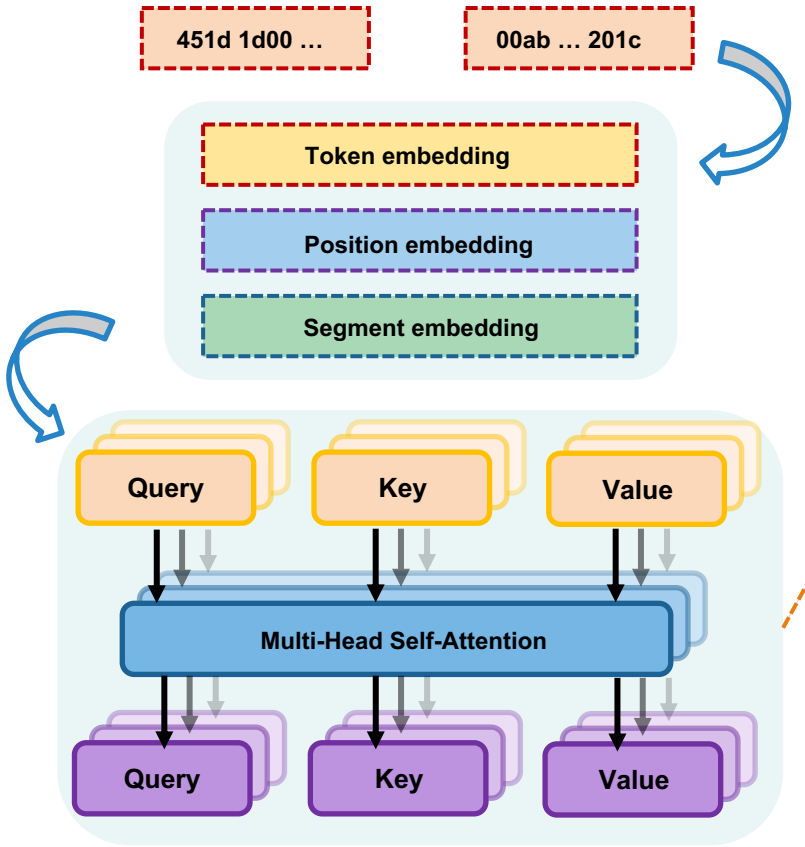
由报文符号组成的流量表示

研究内容：预训练

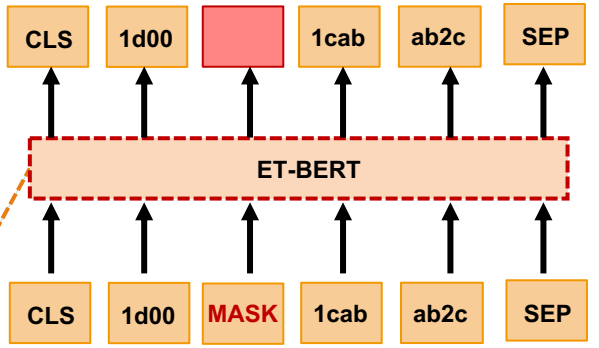
BURST token series

Token2embedding

Pre-training



Masked BURST Model



$$L_{MBM} = -\sum_{i=1}^{k} \log(P(MASK_i = token_i|\bar{X};\theta))$$

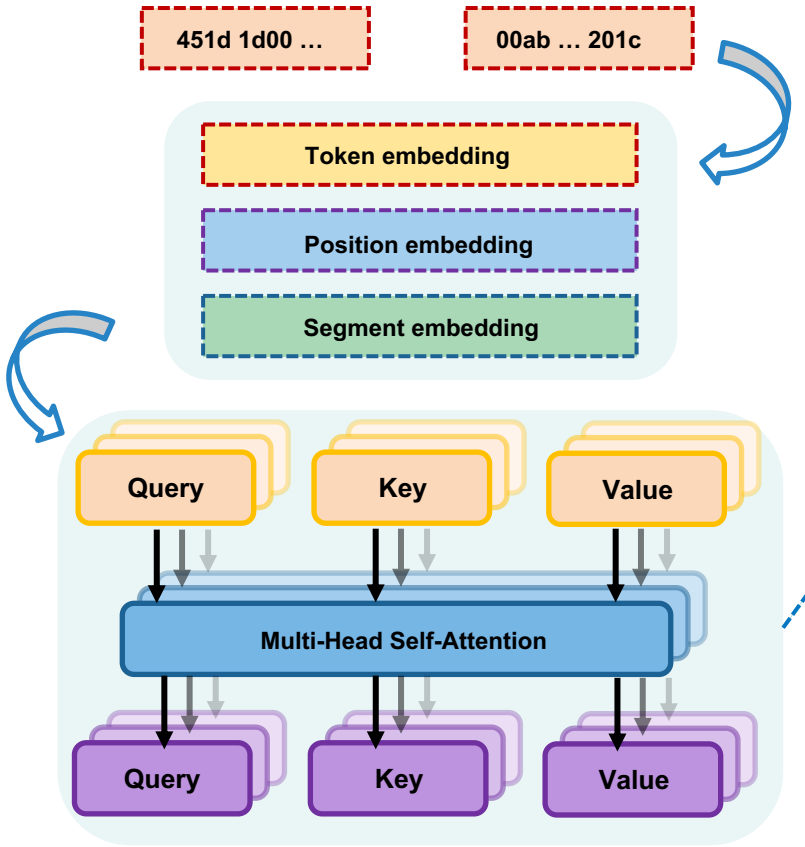
- 基于网络通信架构，短时间相同对端通信的报文信息客观存在紧密的上下文关联
- **BURST掩码**任务的提出是为了在无标注的加密流量报文中挖掘流量报文的共性关联关系

研究内容：预训练

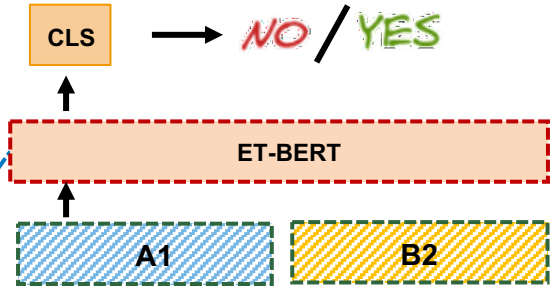
BURST token series

Token2embedding

Pre-training



Same-Origin BURST Prediction



$$L_{SBP} = -\sum_{j=1}^{n} \log(P(y_j|B_j;\theta))$$

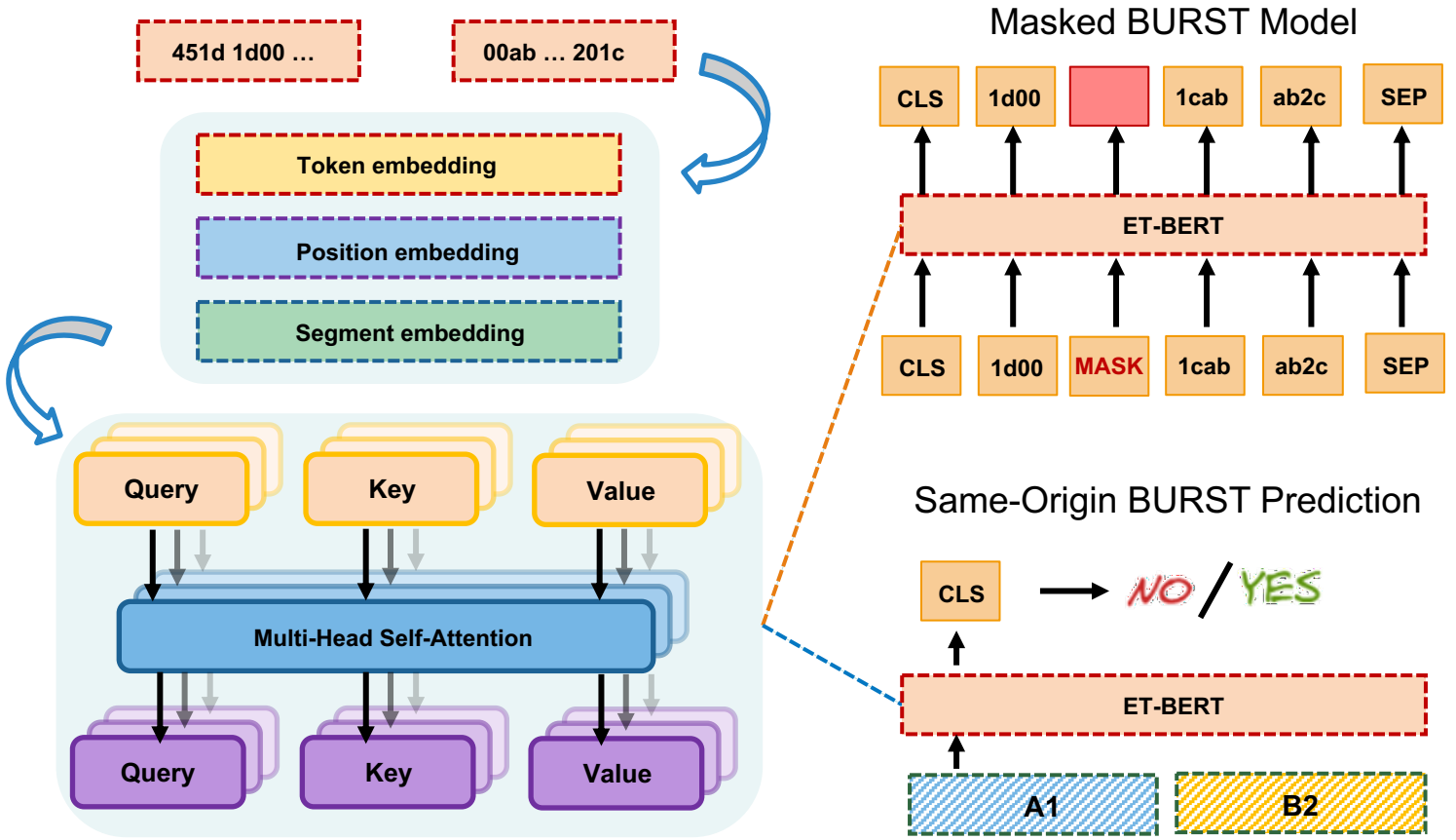
- 基于网络通信架构，短时间相同对端通信的流量客观存在同源关系（协议，IP，端口相同）
- **同源BURST预测**任务的提出是为了在无标注的加密流量报文中挖掘流量的同源关系

研究内容：预训练

BURST token series

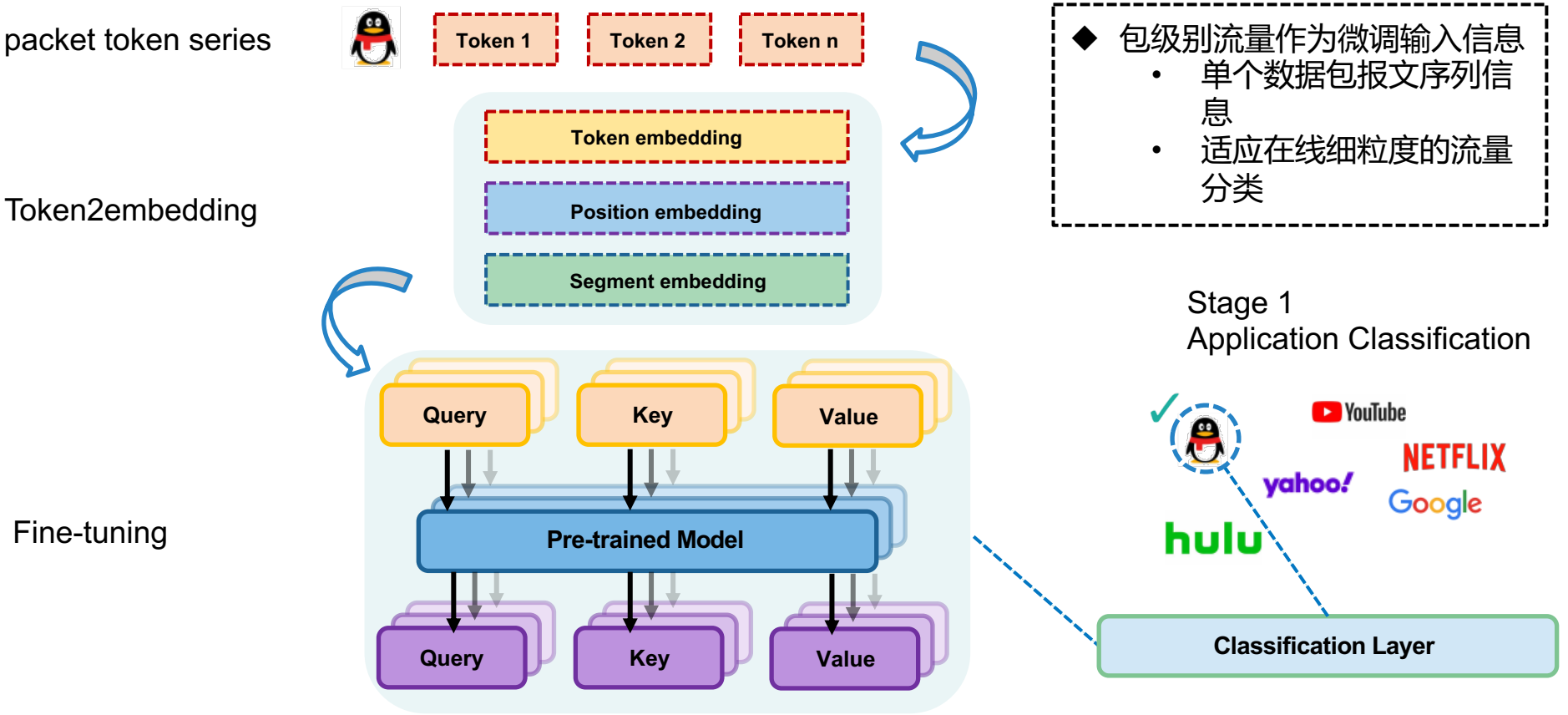
Token2embedding

Pre-training



- **客观现象:** 短时间相同对端通信的报文存在紧密的网络结构与信息关联关系
- 掩码任务适用于挖掘报文内容之间的关联性，同源任务适用于挖掘流量通信结构的同源性

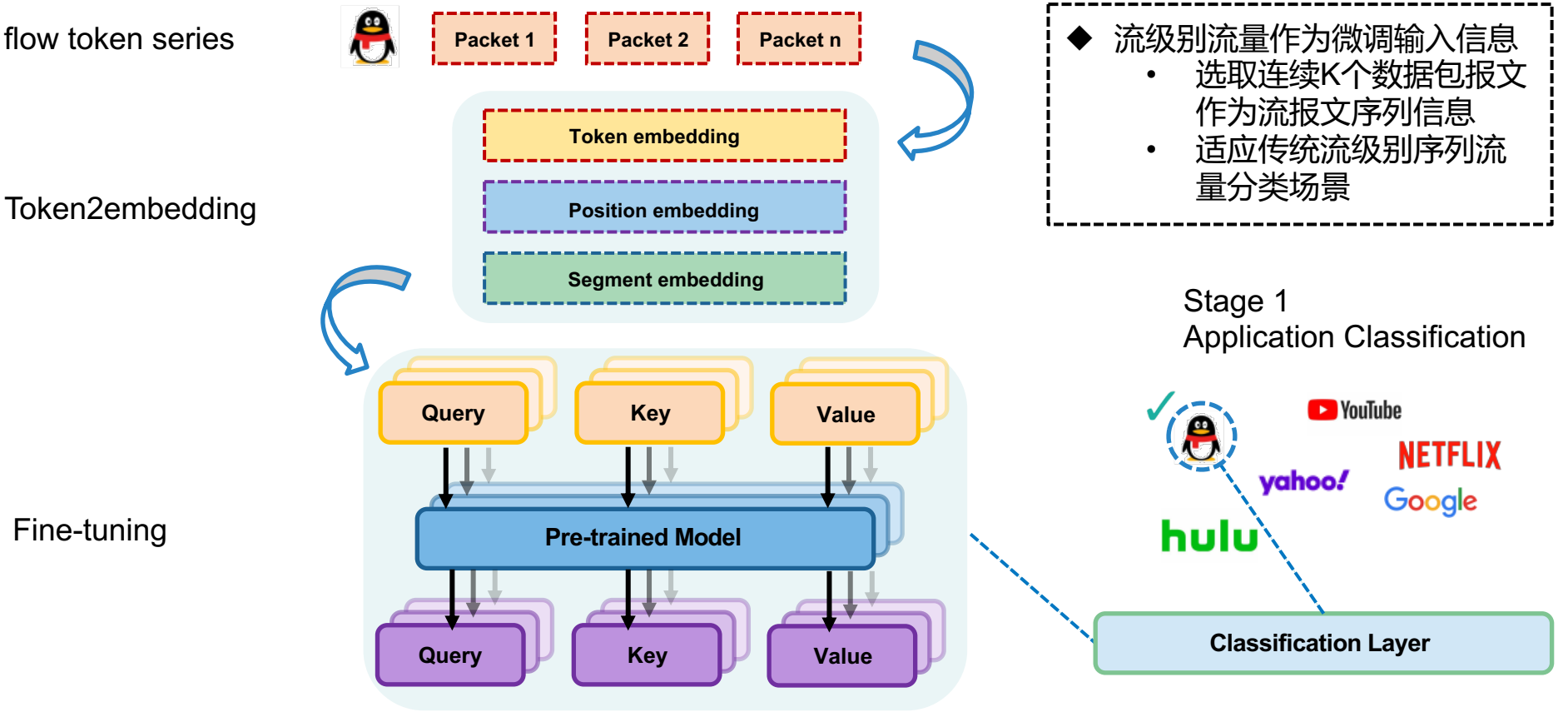
研究内容：微调



◆ 包级别流量作为微调输入信息
- 单个数据包报文序列信息
- 适应在线细粒度的流量分类

➢ 微调建立在预训练好的模型基础，输入带标注的流/包样本对流量表征进行特定场景适应

研究内容：微调



◆ 流级别流量作为微调输入信息
- 选取连续K个数据包报文作为流报文序列信息
- 适应传统流级别序列流量分类场景

➢ 微调建立在预训练好的模型基础，输入带标注的流/包样本对流量表征进行特定场景适应

实验评估

实验评估目标

- ET-BERT 能否适应新型加密协议场景，例如TLS 1.3？
- 模型能否在多种不同加密场景下保持性能稳定，例如VPN，Tor？
- 在不均衡数据场景和小样本场景下，模型是否出现性能滑坡？
- 预训练任务和BURST表示是否真的起到作用？

实验对比方法

- 指纹规则库构建算法 FlowPrint [1]
- 人工特征+机器学习 AppScanner，CUMUL，BIND，K-FP [2,3,4,5]
- 深度学习模型 DF，FS-Net，GraphDApp，TSCRNN，Deeppacket，PERT [6,7,8,9,10,11]

[1] FlowPrint: Semi-Supervised Mobile-App Fingerprinting on Encrypted Network Traffic, NDSS'20
[2] Robust Smartphone App Identification via Encrypted Network Traffic Analysis, TIFS'18
[3] Website Fingerprinting at Internet Scale, NDSS'16
[4] Adaptive Encrypted Traffic Fingerprinting with Bi-Directional Dependence, ACSAC'16
[5] K-fingerprinting: A Robust Scalable Website Fingerprinting Technique, USENIX'16
[6] Deep Fingerprinting: Undermining Website Fingerprinting Defenses with Deep Learning, CCS'18
[7] FS-Net: A Flow Sequence Network For Encrypted Traffic Classification, INFOCOM'19
[8] Accurate Decentralized Application Identification via Encrypted Traffic Analysis Using Graph Neural Networks, TIFS'21
[9] TSCRNN: A Novel Classification Scheme of Encrypted Traffic Based on Flow Spatiotemporal Features for Efficient Management of IIoT, CN'21
[10] Deep Packet: A Novel Approach for Encrypted Traffic Classification Using Deep Learning, SC'20
[11] PERT: Payload Encoding Representation from Transformer for Encrypted Traffic Classification, ITU'20

实验评估

实验数据集

- Cross-Platform，公开数据集，包含了来自美国、印度和中国的196类iOS应用和215类安卓应用。
- USTC-TFC，公开数据集，分别包含20类良性应用和恶意应用。
- ISCX-VPN，公开数据集，包含了17类Open-VPN隧道下的应用。
- ISCX-Tor，公开数据集，包含了16类Tor网络下的应用。
- CSTNET-TLS 1.3，自采数据集（已公开）包含 120类TLS 1.3应用。

Task	Dataset	#Flow	#Packet	#Label
GEAC	Cross-Platform(iOS) [35]	20,858	707,717	196
	Cross-Platform(Android) [35]	27,846	656,044	215
EMC	USTC-TFC [39]	9,853	97,115	20
ETCV	ISCX-VPN-Service [9]	3,694	60,000	12
	ISCX-VPN-App [9]	2,329	77,163	17
EACT	ISCX-Tor [10]	3,021	80,000	16
EAC-1.3	CSTNET-TLS 1.3 (Ours)	46,372	581,709	120

实验评估

Dataset	ISCX-Tor				ISCX-VPN-App				CSTNET-TLS 1.3			
Method	AC	PR	RC	F1	AC	PR	RC	F1	AC	PR	RC	F1
AppScanner[32]	0.6722	0.3756	0.4422	0.3913	0.6266	0.4864	0.5198	0.4935	0.6662	0.6246	0.6327	0.6201
CUMUL[23]	0.6606	0.3850	0.4416	0.3918	0.5365	0.4129	0.4535	0.4236	0.5391	0.4942	0.5060	0.4904
BIND[1]	0.7185	0.4598	0.4515	0.4511	0.6767	0.5152	0.5153	0.4965	0.7964	0.7605	0.7650	0.7560
K-fp[10]	0.6472	0.5576	0.5849	0.5522	0.6070	0.5478	0.5430	0.5303	0.4036	0.3969	0.4044	0.3902
FlowPrint[33]	0.9092	0.3820	0.3661	0.3654	0.8767	0.6697	0.6651	0.6531	0.1261	0.1354	0.1272	0.1116
DF[31]	0.7533	0.6228	0.6010	0.5850	0.6116	0.5706	0.4752	0.4799	0.7936	0.7721	0.7573	0.7602
FS-Net[18]	0.6071	0.5080	0.5350	0.4590	0.6647	0.4819	0.4848	0.4737	0.8639	0.8404	0.8349	0.8322
GraphDApp[29]	0.6836	0.4864	0.4823	0.4488	0.6328	0.5900	0.5472	0.5558	0.7034	0.6464	0.6510	0.6440
TSCRNN[17]	-	0.9490	0.9480	0.9480	-	-	-	-	-	-	-	-
Deeppacket[21]	0.7449	0.7549	0.7399	0.7473	0.9758	0.9785	0.9745	0.9765	0.8019	0.4315	0.2689	0.4022
PERT[11]	0.7682	0.4424	0.4446	0.4345	0.8229	0.7092	0.7173	0.6992	0.8915	0.8846	0.8719	0.8741
ET-BERT(flow)	0.8311	0.5564	0.6448	0.5886	0.8519	0.7508	0.7294	0.7306	0.9510	0.9460	0.9419	0.9426
ET-BERT(packet)	**0.9921**	**0.9923**	**0.9921**	**0.9921**	**0.9962**	**0.9936**	**0.9938**	**0.9937**	**0.9737**	**0.9742**	**0.9742**	**0.9741**

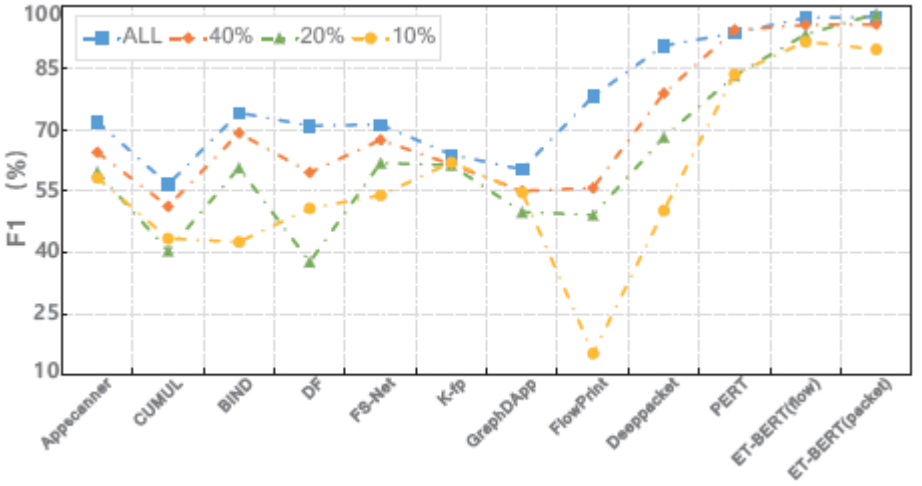
- ET-BERT 在**全加密网络**、**匿名网络**等多种加密场景均达到识别效果的**SOTA**
- 新型加密协议 TLS 1.3、隧道、匿名网络流量
  - ① 协议明文信息减少 — 规则构建和明文依赖缺失 (FlowPrint**性能差**)
  - ② 特征变化复杂 — 基于人为构建特征的局限性暴露 (统计特征方法**性能差**)
  - ③ 场景特征不唯一 — 基于场景设计的序列特征泛化性不足（传统深度学习方法**性能差**）

实验评估

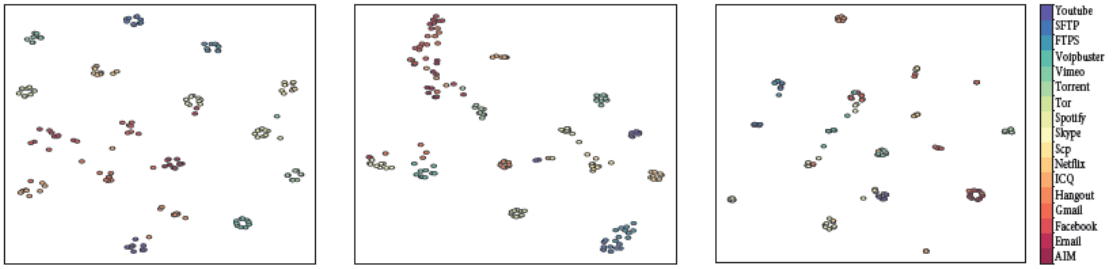
Dataset	Cross-Platform(IOS)				Cross-Platform(Android)				ISCX-VPN-Service				USTC-TFC			
Method	AC	PR	RC	F1	AC	PR	RC	F1	AC	PR	RC	F1	AC	PR	RC	F1
AppScanner[32]	0.3205	0.2103	0.2173	0.2030	0.3868	0.2523	0.2594	0.2440	0.7182	0.7339	0.7225	0.7197	0.8954	0.8984	0.8968	0.8892
CUMUL[23]	0.2910	0.1917	0.2081	0.1875	0.3525	0.2221	0.2409	0.2189	0.5610	0.5883	0.5676	0.5668	0.5675	0.6171	0.5738	0.5513
BIND[1]	0.3770	0.2566	0.2715	0.2484	0.4728	0.3126	0.3253	0.3026	0.7534	0.7583	0.7488	0.7420	0.8457	0.8681	0.8382	0.8396
K-fp[10]	0.2155	0.2037	0.2069	0.2003	0.2248	0.2113	0.2104	0.2052	0.6430	0.6492	0.6417	0.6395	-	-	-	-
FlowPrint[33]	0.9254	0.9438	0.9254	0.9260	0.8698	0.9007	0.8698	0.8702	0.7962	0.8042	0.7812	0.7820	0.8146	0.6434	0.7002	0.6573
DF[31]	0.3106	0.2232	0.2179	0.2140	0.3862	0.2595	0.2620	0.2527	0.7154	0.7192	0.7104	0.7102	0.7787	0.7883	0.7819	0.7593
FS-Net[18]	0.3712	0.2845	0.2754	0.2655	0.4846	0.3544	0.3365	0.3343	0.7205	0.7502	0.7238	0.7131	0.8846	0.8846	0.8920	0.8840
GraphDApp[29]	0.3245	0.2450	0.2392	0.2297	0.4031	0.2842	0.2786	0.2703	0.5977	0.6045	0.6220	0.6036	0.8789	0.8226	0.8260	0.8234
TSCRNN[17]	-	-	-	-	-	-	-	-	-	0.9270	0.9260	0.9260	-	0.9870	0.9860	0.9870
Deeppacket[21]	0.9204	0.8963	0.8872	0.9034	0.8805	0.8004	0.7567	0.8138	0.9329	0.9377	0.9306	0.9321	0.9640	0.9650	0.9631	0.9641
PERT[11]	0.9789	0.9621	0.9611	0.9584	0.9772	0.8628	0.8591	0.8550	0.9352	0.9400	0.9349	0.9368	0.9909	0.9911	0.9910	0.9911
ET-BERT(flow)	**0.9844**	0.9701	0.9632	0.9643	**0.9865**	0.9324	**0.9266**	**0.9246**	0.9729	0.9756	0.9731	0.9733	**0.9929**	**0.9930**	**0.9930**	**0.9930**
ET-BERT(packet)	0.9810	**0.9757**	**0.9772**	**0.9754**	0.9728	**0.9439**	0.9119	0.9206	**0.9890**	**0.9891**	**0.9890**	**0.9890**	0.9915	0.9915	0.9916	0.9916

- ET-BERT 在**种类规模大**、**样本分布不均衡**、**恶意服务**等场景均取得识别效果的**SOTA**
- 大规模应用类别
  - ① 应用区分难度增大 — 强特征的指纹构建 (**无法适应**未来加密网络环境)
  - ② 应用区分难度增大 — 特征选择与构建的局限性(**无法适应**多种场景和高速网络)
- 数据不均衡分布
  - ① 少样本应用的标签漂移问题

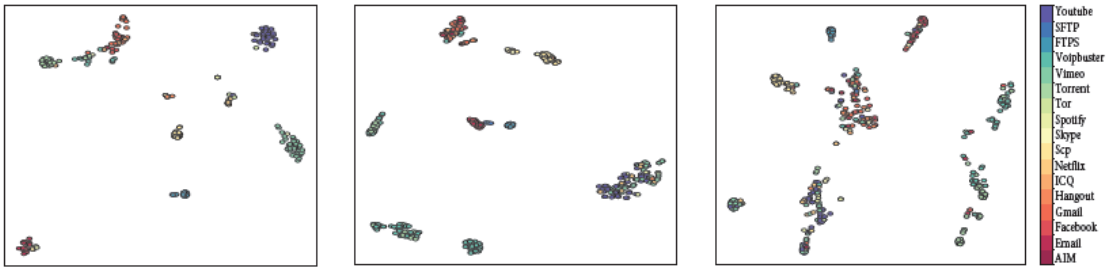
实验评估



- 小样本测试对比
① ET-BERT在所有方法对比中均取得最好的识别结果
② 样本数据规模降低到10%，ET-BERT性能波动最稳定



(a) Representation with ET-BERT (b) Representation with Transformer at packet (c) Representation with Deeppacket



(e) Representation with PERT (f) Representation with Transformer at flow (g) Representation with DF

- 不均衡场景的结果可视化对比
① ET-BERT对所有类别的数据建立了更清晰的边界

消融分析

	Method	SBP	MBM	PT-P	PT-B	FT-f	FT-cf	FT-P	AC	PR	RC	F1
	ET-BERT(packet)(full model)	✓	✓	×	✓	×	×	✓	**0.9471**	**0.9462**	**0.9412**	**0.9395**
1	w/o SBP	×	✓	×	✓	×	×	✓	0.9000	0.9142	0.9000	0.8998
2	w/o MBM	✓	×	×	✓	×	×	✓	0.8471	0.8666	0.8471	0.8462
3	w/o BURST	✓	✓	✓	×	×	×	✓	0.9235	0.9386	0.9235	0.9258
4	ET-BERT(flow)	✓	✓	×	✓	✓	×	×	0.8133	0.7661	0.7374	0.7387
5	concatenated-flow(cf)	✓	✓	×	✓	×	✓	×	0.8229	0.7488	0.6812	0.6961
6	w/o pre-training(packet)	×	×	×	×	×	×	✓	0.5882	0.6152	0.5882	0.5638

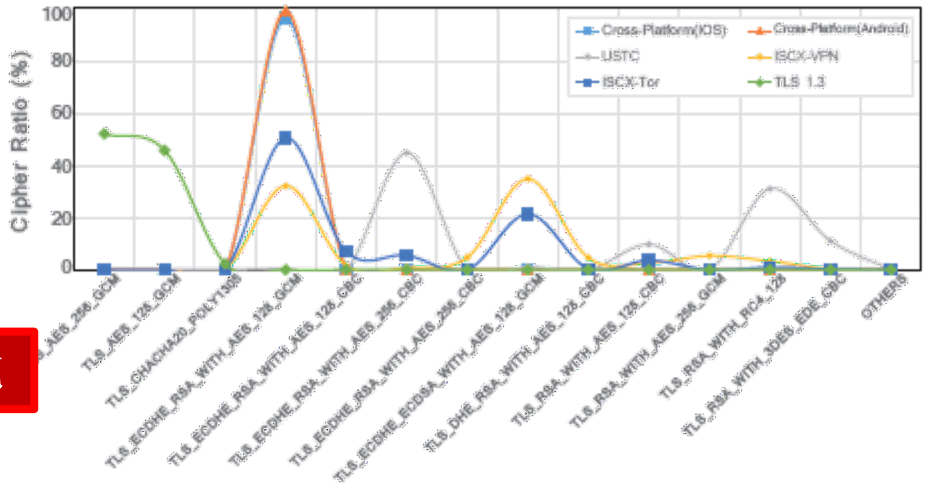
- **自监督任务**：掩码任务和同源任务对于预训练阶段都有增益的贡献，其中掩码任务的增益更为显著。
- **BURST结构**：预训练阶段的流量结构采用BURST而非数据包，对预训练阶段也有增益贡献，但相比于自监督任务并不显著。
- **微调**：作为微调阶段的流量结构，数据包比任意形式的数据流都有更好的贡献。
- **预训练**：预训练阶段对模型的泛化性具有决定性影响。

开放性问题

- **加密流量分析的可行性**
  - 不同密码套件应用带来的随机性差异
  - 报文包头域的关联与差异
- **场景适应**
  - 流量的动态持续性变化将带来场景样本的改变，泛化性的稳定
  - 预训练过程引入的噪声信息或虚假关联，确定因果关系与降噪处理
- **模型投毒与对抗检测**
  - 预训练模型面临毒化与对抗的安全隐患，在预训练阶段加入对抗攻击手段

性能持续性稳定探索以及潜在安全风险对抗

Ciphers/Tests	AES(GCM)	AES(CBC)	CHA20	ARC4	3DES
Monobit	0.7918	0.2585	0.9761	0.5687	0.4099
Block Frequency	0.6316	0.0791	0.0176	0.4821	0.6434
Independent Runs	0.8824	0.1672	0.8966	0.7052	0.4241
Longest Runs	0.7198	0.3148	0.5134	0.5156	0.2889
Spectral	0.6202	0.9707	0.9415	0.6729	0.5756
Overlapping Patterns(OP)	0.0519	0.9856	0.1002	0.9089	0.4762
Non OP	0.8148	0.1967	0.4445	0.0096	0.5156
Universal	0.8501	0.3277	0.1149	0.0416	0.3062
Serial	0.7690	0.4539	0.1600	0.6068	0.8381
Approximate Entropy	0.9239	0.5226	0.3371	0.3470	0.3611
Cumulative Sums	0.9496	0.4512	0.7355	0.1742	0.4043
Random Excursions(RE)	0.1811	0.1232	0.4112	0.9424	0.9091
RE Variant	0.4805	0.0119	0.9542	0.5978	0.9065
Matrix Rank	0.5674	0.4890	0.0880	0.0504	0.1447
Linear Complexity	0.6235	0.4519	0.7428	0.0952	0.9384



总结

- 提出加密流量预训练方法 **ET-BERT**
  - ① 适用于多种场景任务的通用加密流量预训练架构
  - ② 不依赖大量标注数据，在无标注流量中预训练并快速迁移到特定任务
  - ③ 2种自监督任务能够有效捕捉流量不同层级的关联关系
- **ET-BERT**展现出强大的性能
  - ① 在6种不同场景下对比11种现有方法，取得最好的加密流量识别效果
  - ② 面对全加密趋势的新型加密流量依旧保持最佳的识别效果
  - ③ 在小样本、不均衡等场景下流量识别性能下降最少
- 未来工作
  - ① 预训练模型的迁移应用的安全性保障，例如迁移过程的数据投毒和模型后门
  - ② 在线流量处理所需要的轻量级预训练模型，例如模型压缩与剪枝

论文与代码

ET-BERT

codebeat B | license MIT | arXiv 1909.05658

The repository of ET-BERT, a network traffic classification model on encrypted traffic.

ET-BERT is a method for learning datagram contextual relationships from encrypted traffic, which could be **directly applied to different encrypted traffic scenarios and accurately identify classes of traffic.** First, ET-BERT employs multi-layer attention in large scale unlabelled traffic to learn both inter-datagram contextual and inter-traffic transport relationships. Second, ET-BERT could be applied to a specific scenario to identify traffic types by fine-tuning the labeled encrypted traffic on a small scale.

ET-BERT: A Contextualized Datagram Representation with Pre-training Transformers for Encrypted Traffic Classification

Xinjie Lin, Gang Xiong, Gaopeng Gou, Zhen Li, Junzheng Shi, Jing Yu

Encrypted traffic classification requires discriminative and robust traffic representation captured from content-invisible and imbalanced traffic data for accurate classification, which is challenging but indispensable to achieve network security and network management. The major limitation of existing solutions is that they highly rely on the deep features, which are overly dependent on data size and hard to generalize on unseen data. How to leverage the open-domain unlabeled traffic data to learn representation with strong generalization ability remains a key challenge. In this paper,we propose a new traffic representation model called Encrypted Traffic Bidirectional Encoder Representations from Transformer (ET-BERT), which pre-trains deep contextualized datagram-level representation from large-scale unlabeled data. The pre-trained model can be fine-tuned on a small number of task-specific labeled data and achieves state-of-the-art performance across five encrypted traffic classification tasks, remarkably pushing the F1 of ISCX-Tor to 99.2% (4.4% absolute improvement), ISCX-VPN-Service to 98.9% (5.2% absolute improvement), Cross-Platform (Android) to 92.5% (5.4% absolute improvement), CSTNET-TLS 1.3 to 97.4% (10.0% absolute improvement). Notably, we provide explanation of the empirically powerful pre-training model by analyzing the randomness of ciphers. It gives us insights in understanding the boundary of classification ability over encrypted traffic. The code is available at: this https URL.

Comments: This work has been accepted in Security, Privacy, and Trust track at The Web Conference 2022 (WWW'22)(see this https URL)
Subjects: **Cryptography and Security (cs.CR)**; Artificial Intelligence (cs.AI); Networking and Internet Architecture (cs.NI)

Thanks! Q&A

Homepage

Jing Yu

Email: yujing02@iie.ac.cn

Homepage: https://mmlab-iie.github.io/

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

中国科学院大学
University of Chinese Academy of Sciences