SIGCOMM 2025

The Sweet Danger of Sugar: Debunking Representation Learning for Encrypted Traffic Classification

汇报人：唐臻宇

目录
content

1 Background and Motivation

2 Approach

3 Experiment

4 Thinking and Inspiration

Represention Learning

基于表示学习的模型（如BERT、MAE）在NLP、CV中取得了巨大成功，近期越来越多研究将这些方法应用在加密流量分类任务，提出了多种预训练的表示模型如下，并取得了不错的效果

Model	Pre-training				Downstream Classification			
	Architecture	Embedding Size	Task Types	Dataset	Cleaning	Split	# Tasks	Datasets
PacRep [33]	BERT	768	None	Not needed	Partial	Packet	6	A, B, +
PERT [18]	ALBERT	768	MAE	≠	No	Flow	2	A, +
ET-BERT [27]	BERT	768	MAE, SBP	∩	Partial	Packet	7	A, B, C, +
PTU [42]	BERT	768	MAE, SSP, HIP, FIP	≠	No	Packet	7	A, B, C, +
TrafficFormer [54]	BERT	768	MAE, SODF	∩	Partial	Packet	6	A, B, C, +
netFound [17]	BERT	1024	MAE	≠	Partial	Flow	5	A, +
YaTC [53]	ViT	192	MAE	=	No	Unknown	4	A, B, +
NetMamba [49]	Mamba	256	MAE	=	Partial	Flow	6	A, B, +
Pcap-Encoder	T5	768	Autoencoder, Q&A	≠	Full	Flow	6	A, B, C

Datasets: A=ISCX-VPN, B=USTC-TFC, C=CSTNET-TLS1.3, +=other

Table 1: Summary of representation learning models for traffic classification. Pitfalls are highlighted in red.

Motivation

The Sweet of Danger

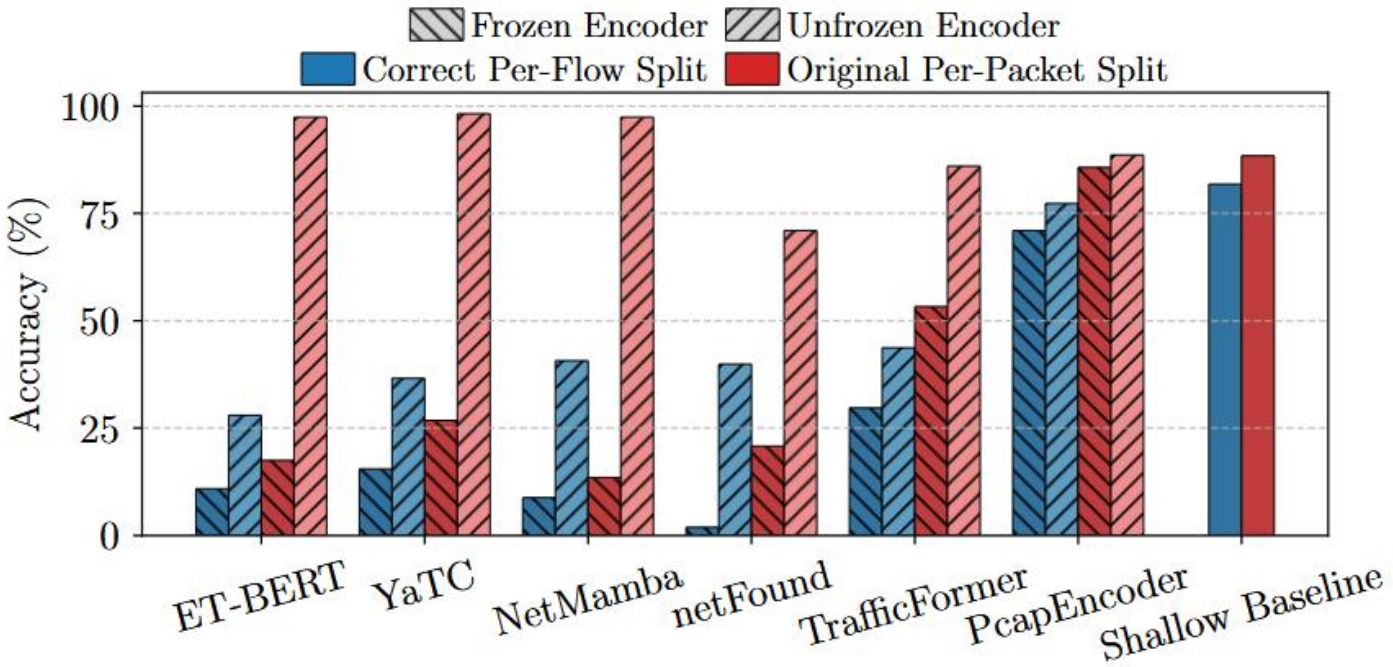


Figure 1: Accuracy of classifiers evaluated (TLS-120 dataset, packet classification task). Models' performance collapses when properly tested. *Pcap-Encoder* is the only model that maintains good performance. However, a simple shallow baseline surpass all representation learning-based methods.

Motivation

The Sweet of Danger

problems

- 模型采用per-flow split后，模型性能显著下降——模型是否真的学到了有意义的表示，还是仅仅利用了数据集中的捷径?
- 在微调阶段冻结编码器，模型性能大幅下降——加密流量中是否存在可学习的语义信息?
- 许多方法在数据处理过程中未能充分保障数据的完整性与一致性

fingdings

- per-packet split 导致严重的数据泄露，使模型倾向于捷径学习，如流标识符
- 直接从加密载荷（payload）中学习有效模式极为困难
- 先前研究采用不一致的数据清洗方法，导致模型在实际部署中出现显著性能偏差

approach

- 使用 per-flow split 替代 per-packet split，以杜绝数据泄露
- 专注于从数据包标头中提取有效特征，而非依赖加密字段
- 采用合理且一致的数据清洗方式，保证数据的完整性及结果的可比性

Contribution

- **系统性批判现有方法：**指出多数现有工作在数据划分（per-packet split）、预训练任务设计、评估方式上存在严重缺陷。
- **提出新的评估框架：**强调应使用 per-flow split 和 frozen encoder 来检验表示学习的有效性。
- **提出新模型：**Pcap-Encoder：基于T5架构，专注于从协议头部提取语义信息，忽略加密载荷。
- **公开代码与数据集：**提供可复现的基准测试框架，促进社区健康发展。

目录
content

1 Background and motivation

2 Approach

3 Experiment

4 Thinking and inspiration

Approach

Data cleaning

- **过滤无关协议**：移除与分类任务无关的协议（如ARP、DHCP、广播协议等）——这些协议通常与目标任务（如应用识别、恶意流量分类）无关，若保留可能干扰模型训练或引入噪声。

- **避免基于最小尺寸的过滤**：不根据包大小或流长度进行过滤（例如不移除80字节的包或短流）。——这类过滤会改变数据分布，可能丢失重要信号（如TCP握手包、确认包），影响模型泛化能力。对于先前工作（如ET-BERT、TrafficFormer）会过滤小包或短流，作者认为这种做法不合理

- **保证类别分布完整性**：在测试集中不进行类别平衡或过采样/欠采样，保持原始数据分布；仅在训练集可适当进行类别平衡——测试集应反映真实场景中的类别分布，避免因人为平衡导致性能评估失真，训练时可采用平衡采样来处理类别不平衡

- **头部信息匿名化或保留**：在预训练阶段保留所有头部信息，在下游任务中可根据需要选择是否匿名化（如随机化IP、端口）——预训练需充分利用头部结构信息学习通用表示；下游任务匿名化可迫使模型学习更泛化的特征，避免记忆特定IP或端口

Approach

Dataset Splitting

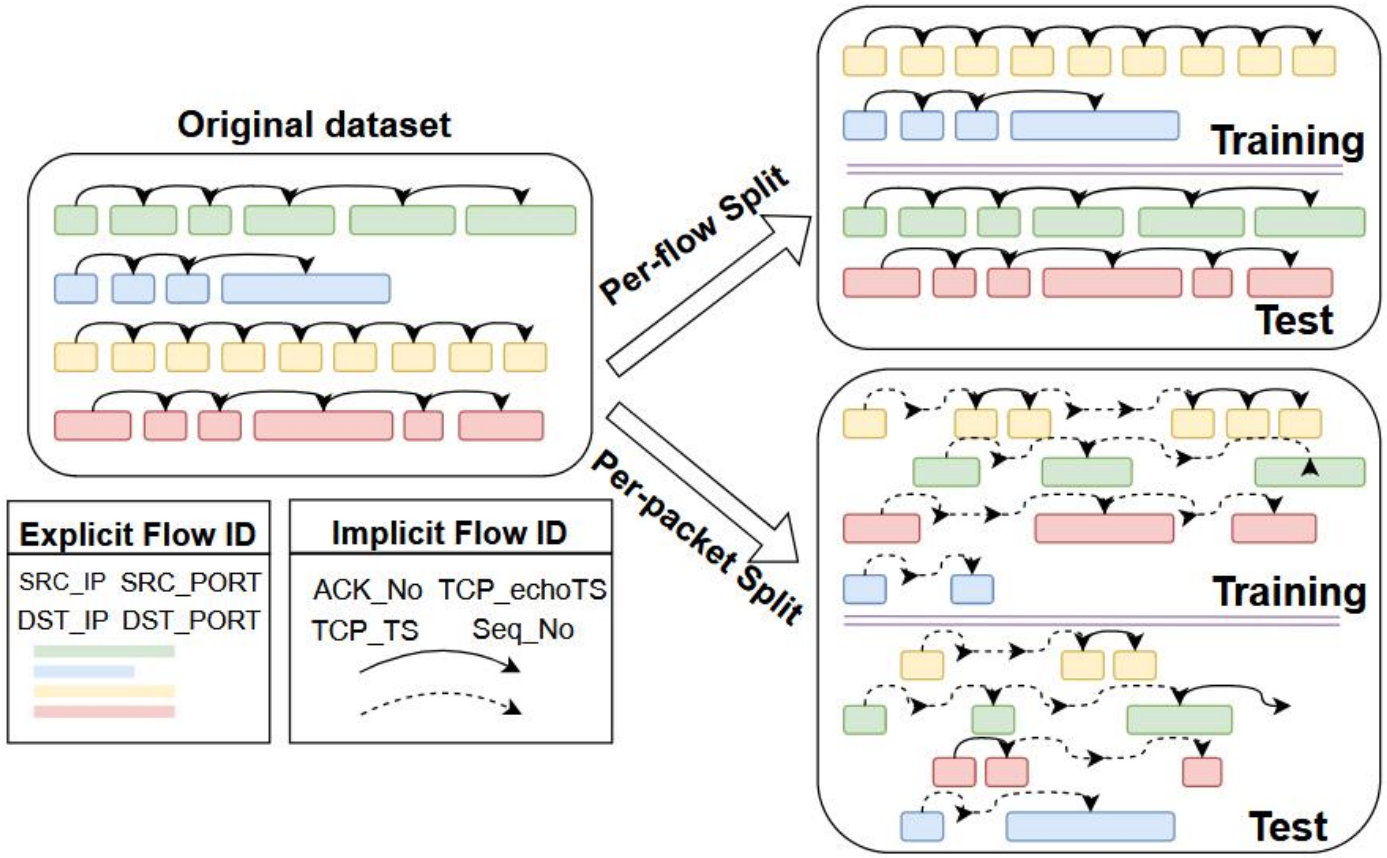
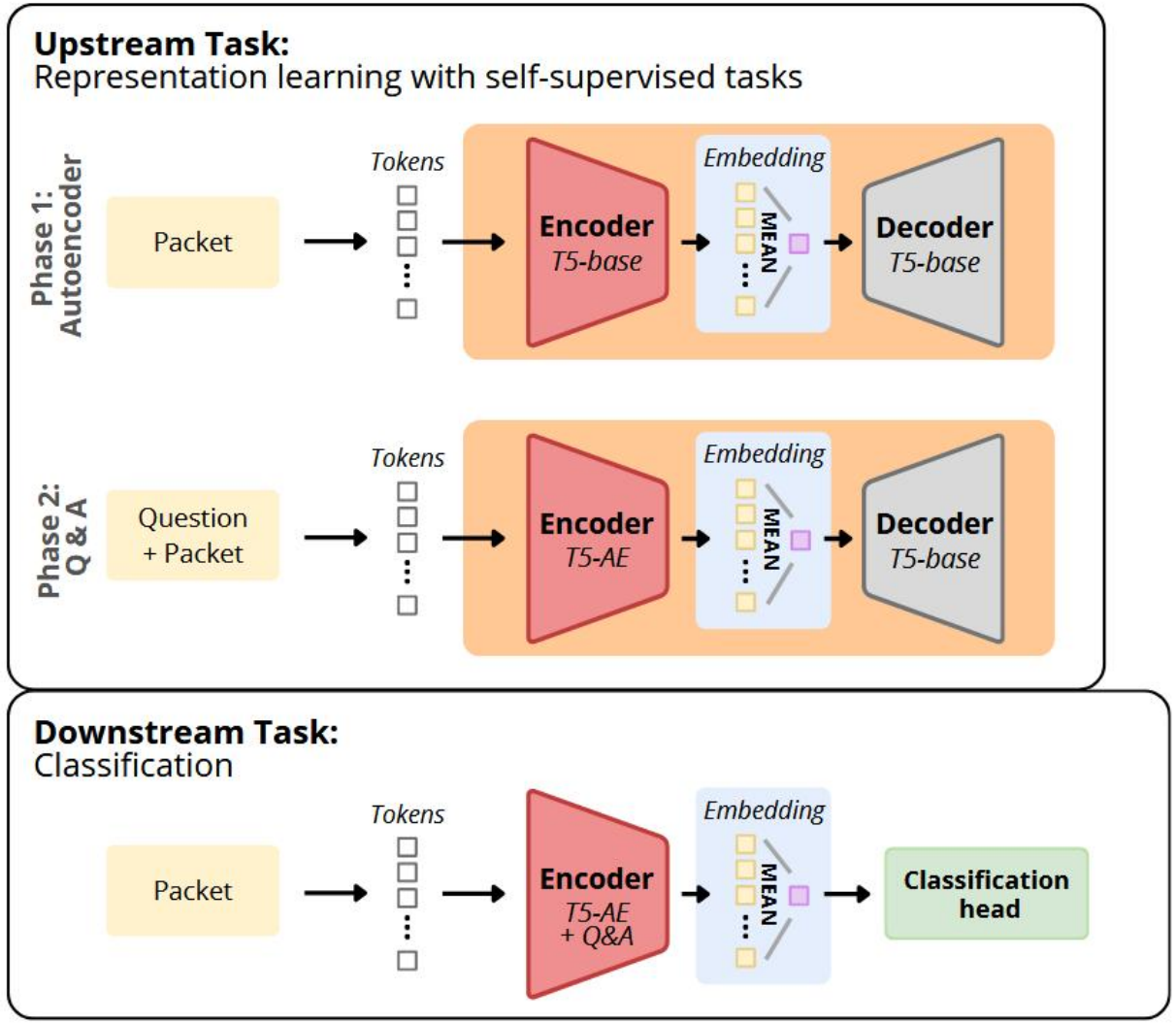


Figure 3: Per-flow and per-packet split.

Approach

Pcap-Encoder



核心任务

利用双阶段训练，捕获字节之间的上下文关系，并自动提取某些数据包标头字段的语义

Phase1:自编码器重建

基于MAE学习从原始字节流中提取有用表示。

Phase2:问答任务(Q&A)

训练模型回答关于协议头部的问题（如IP地址、TTL、校验和等），强化语义理解

目录
content

1 Background and motivation

2 Approach

3 Experiment

4 Thinking and inspiration

Main Results

Traffic Classification Task

Model (Per-flow split)	VPN-binary (2)		VPN-service (6)		VPN-app (16)		USTC-binary (2)		USTC-app (20)		TLS-120	
	AC	F1	AC	F1	AC	F1	AC	F1	AC	F1	AC	F1
ET-BERT	84.7	84.6	71.7	64.2	59.2	43.7	100.0	100.0	84.9	79.6	10.9	6.7
YaTC	83.9	83.9	69.2	60.1	60.9	44.3	99.5	99.5	85.2	78.0	15.5	9.6
NetMamba	75.0	74.5	56.9	49.0	39.6	28.4	97.6	97.5	72.5	57.7	8.8	4.5
TrafficFormer	90.9	90.9	76.5	69.4	67.7	54.4	100.0	100.0	72.0	65.0	29.7	24.0
netFound	76.0	61.9	47.3	36.5	32.9	15.3	99.4	99.4	58.0	30.7	1.9	0.5
Pcap-Encoder	**99.9**	**99.9**	**92.1**	**89.8**	**83.5**	**71.0**	**100.0**	**100.0**	**91.0**	**87.1**	**71.0**	**63.7**

Table 3: Results of *Pcap-Encoder* and the three SoA models for packet classification. Per-flow split, Frozen encoders. We report accuracy (AC) and macro F1-Score (F1). Results below 50% are highlighted in red, best in bold.

Model (Per-flow split)	VPN-app (16)				TLS-120			
	Frozen		Unfrozen		Frozen		Unfrozen	
	AC	F1	AC	F1	AC	F1	AC	F1
ET-BERT	59.2	43.7	82.8	69.7	10.9	6.7	28.0	21.5
YaTC	60.9	44.3	79.1	65.2	15.5	9.6	36.6	31.4
NetMamba	39.6	28.4	80.4	65.9	8.8	4.5	40.7	35.3
TrafficFormer	67.7	54.4	73.1	61.0	29.7	24.0	43.7	38.9
netFound	32.9	15.3	70.4	57.3	1.9	0.5	39.9	35.2
Pcap-Encoder	**83.5**	**71.0**	**85.6**	**74.8**	**71.0**	**63.7**	**77.3**	**69.2**

Table 4: Per-flow split, frozen and unfrozen encoder. Results improve, but models still struggle in challenging setups.

Model (Per-packet split)	VPN-app (16)				TLS-120			
	Frozen		Unfrozen		Frozen		Unfrozen	
	AC	F1	AC	F1	AC	F1	AC	F1
ET-BERT	69.5	64.7	96.8	97.0	17.5	10.2	97.4	96.8
YaTC	73.2	67.7	**98.5**	**98.5**	26.8	17.7	**98.2**	**97.7**
NetMamba	53.5	45.1	98.4	98.4	13.5	5.3	97.4	96.8
TrafficFormer	87.5	85.6	95.6	95.2	53.3	48.2	86.0	83.3
netFound	35.3	18.7	89.6	89.0	10.8	2.3	71.0	67.4
Pcap-Encoder	**91.9**	**90.6**	94.3	93.9	**85.7**	**81.0**	88.6	80.3

Table 5: Per-packet split scenario. Eventually, in this wrong settings and unfrozen encoder, performance reaches the promised > 90% accuracy.

Main Results

ET-BERT Analysis

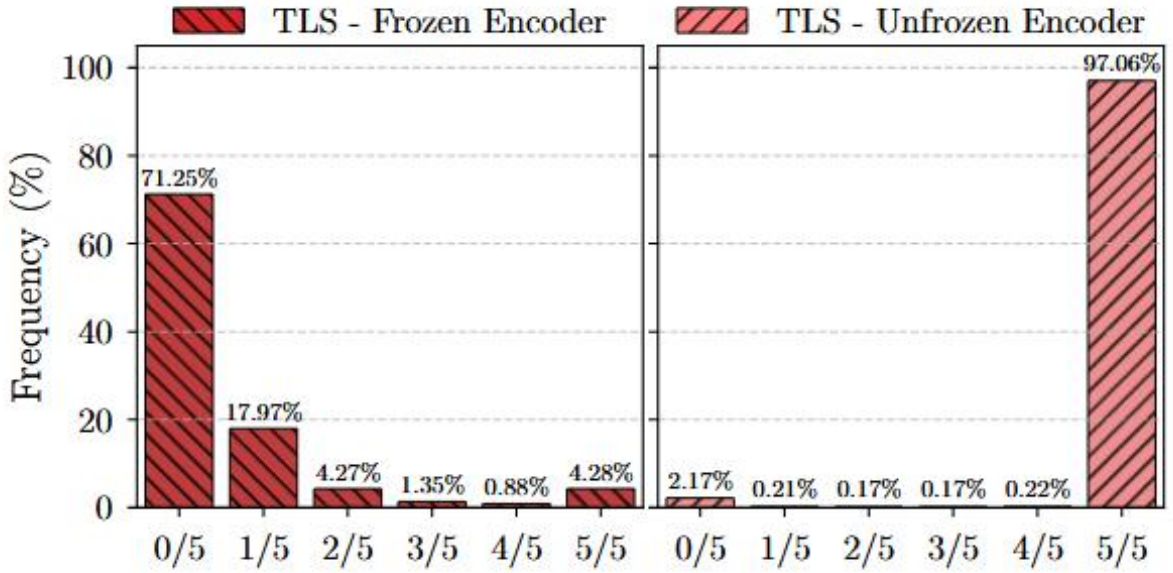


Figure 4: 5-NN purity of embeddings for *ET-BERT*. With a frozen encoder, 71% of points do not have a sample of the same class as TOP-5 neighbour. Situation changes only when the encoder is unfrozen.

Scenario	Dataset	AC	F1
Per-packet Split	Original	97.4	96.8
	w/o SeqNo/AckNo w/o Timestamp (only test)	19.5	15.4
	w/o SeqNo/AckNo w/o Timestamp (train + test)	52.2	48.2
	w/o Pre-training	97.1	96.4
Per-flow Split	Original	28.0	21.5

Table 6: Impact of implicit flow ID on the unfrozen *ET-BERT* and ablation study on the pre-training strategy.

结论

- 错误的数据划分（Per-packet split）导致数据泄露
- 模型在微调中学习了隐式流标识符（Shortcuts）
- 预训练本身是无效的，模型相当于从零开始学习

Ablation Analysis

Model (Per-flow split)	VPN-app (16)	TLS-120
w/o IP addr.	52.5	13.0
w/o header	16.4	1.5
w/o payload	66.7	63.6
base	71.0	63.7

Table 7: Ablation Study on *Pcap-Encoder* in the flow-based split scenario when removing the IPs, headers and payloads (Macro F1-Scores).

有效信息来自IP地址以及协议头，加密载荷payload并不能提供很多有用信息

Model (Per-flow split)	VPN-app (16)		TLS-120	
	base	w/o IP addr	base	w/o IP addr
RF	81.1	72.4	78.0	39.4
XGBoost	82.1	73.2	82.0	41.3
LightGBM	82.6	74.5	82.4	40.6
MLP	65.1	52.5	68.8	30.5

Table 8: Macro F1-Scores of ML baselines. We try the baseline with and without the IP information.

经过精心特征工程的简单模型，其表现可以超越复杂的表示学习模型。这进一步质疑了在流量分类任务中引入沉重表示学习模型的必要性和性价比。

Results

Model Efficiency

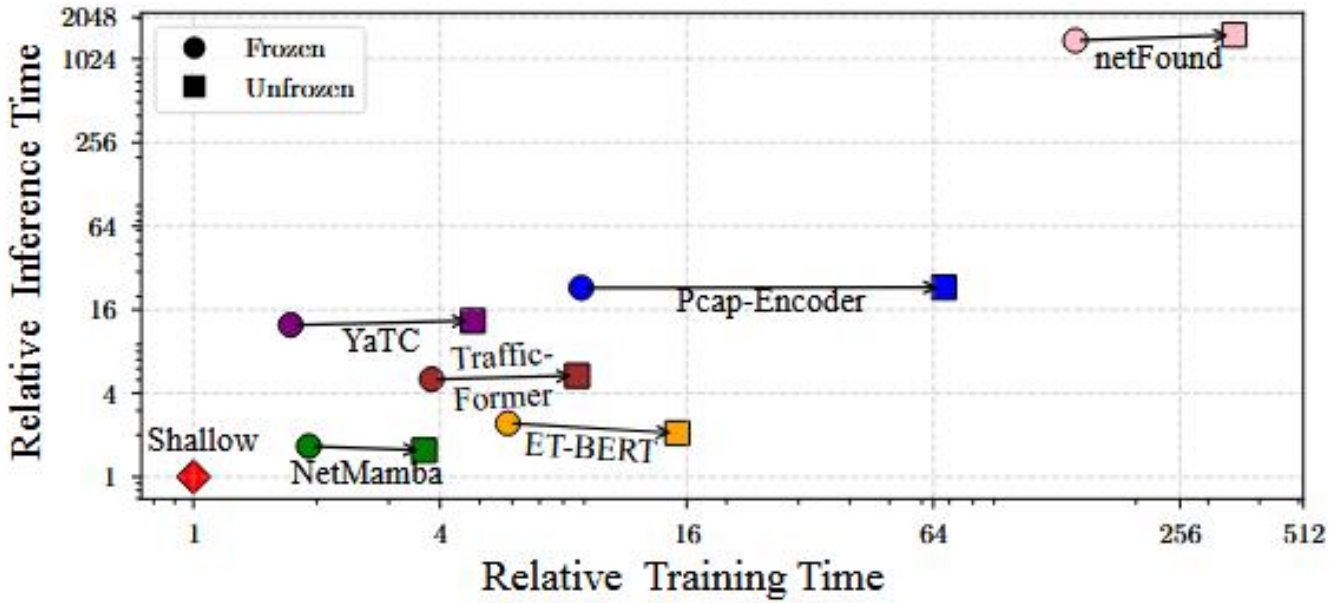


Figure 6: Relative training and inference times. All models are much slower than the shallow baseline, with larger models having the worst ratio (up to 2048× slower at inference).

表示学习模型带来了巨大的计算成本，但其性能优势相对于简单的基线模型并不明显，这使其在实际部署中的实用性受到严重质疑

目录
content

1 Background and motivation

2 Approach

3 Experiment

4 Thinking and inspiration

Thinking and inspiration

The Sweet of Danger

警惕“糖衣陷阱”：过于完美的结果往往隐藏着方法或数据上的缺陷

表示学习的实用性存疑：在当前加密流量分类任务中，复杂表示学习模型未必比传统方法更有效。这也恰恰说明了目前还没有找到一个有效的表示学习方案

使用per-flow split, 测试frozen encoder 下的表示质量，协议头部信息远比加密载荷更有价值

Thank you for your time