Yet Another Traffic Classifier: A Masked Autoencoder Based Traffic Transformer with Multi-Level Flow Representation

論文復現與報告

第29組:

111550018 黃羽良

111550046 周哲煒

動機:
流量辨識能幫助偵測威脅，現有幾類辨識方式有限制

Rule-based method:
檢查基礎特徵(protocol, port num ...)，
已難以準確預測現代複雜網路環境下流量

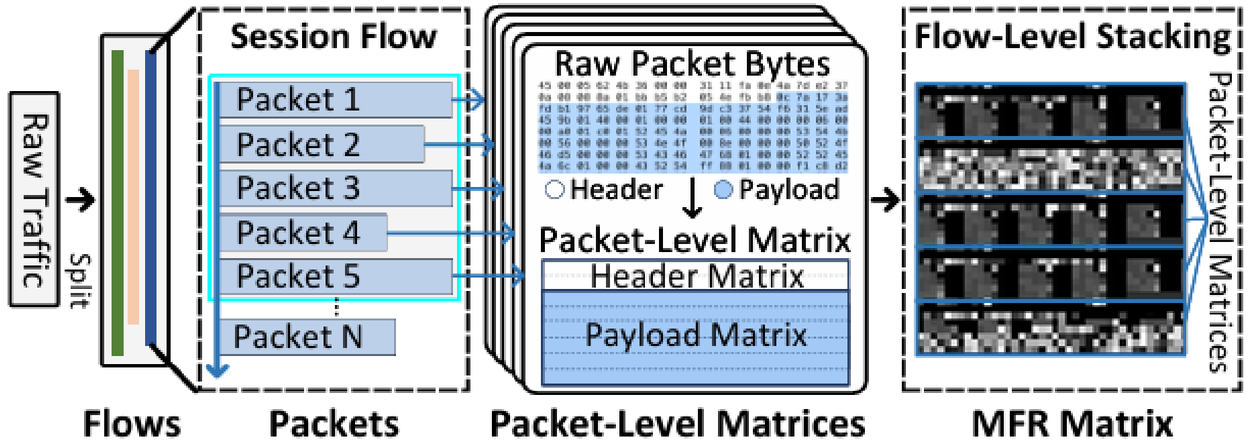
ML-based method:
選擇optimal statistical features為SVM inputs，
依賴專家統計特徵，針對不同場景須選不同特徵

一般DL-based method:

1. 僅以Raw byte表示資料
2. 無traffic characteristic
3. 訓練針對特定情境的classifier要大量時間去label data

YaTC改善部分:

1. Multi-level flow representation (MFR)
   具備封包和位元組關係
2. Masked autoencoder (MAE)



Pre-training method:
NLP: 如BERT，traffic bytes缺少明顯的high-level semantic units
Traditional CV: 裁切、縮放會破壞結構如裁切header
MAE: 隨機遮住圖像的一部分，讓模型去預測被遮住的部分，
位置不變。

MFR building method:
每個封包固定佔用 2 行 Header、6 行 Payload，共40行。
這確保Header不會被Payload 淹沒，且5個封包都有考慮到。
移除Ethernet headers，隨機化IP，但保留方向性。

Traffic transformer:

1.Embedding:
切割MFR為N個不重疊patches

$$x_0 = [x_p^1 E; x_p^2 E; \ldots; x_p^N E] + E_{pos}$$

$x_p^i E$: 2D 的 Patch 映射到 D 維的向量空間 (D = 192)

x_0 : Transformer encoder的初始input

2.Packet-level Attention
產生Query, Key, Value 向量

$$Q = x^l W_O, \quad K = x^l W_K, \quad V = x^l W_V$$

$$Attn(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{D_k}}\right) V$$

複雜度由O(N^2)轉為O(N^2/M) N: patches, M: packets

3.Flow-level Attention

x_r = Pooling (x_p’)，將Patch壓縮成row patch

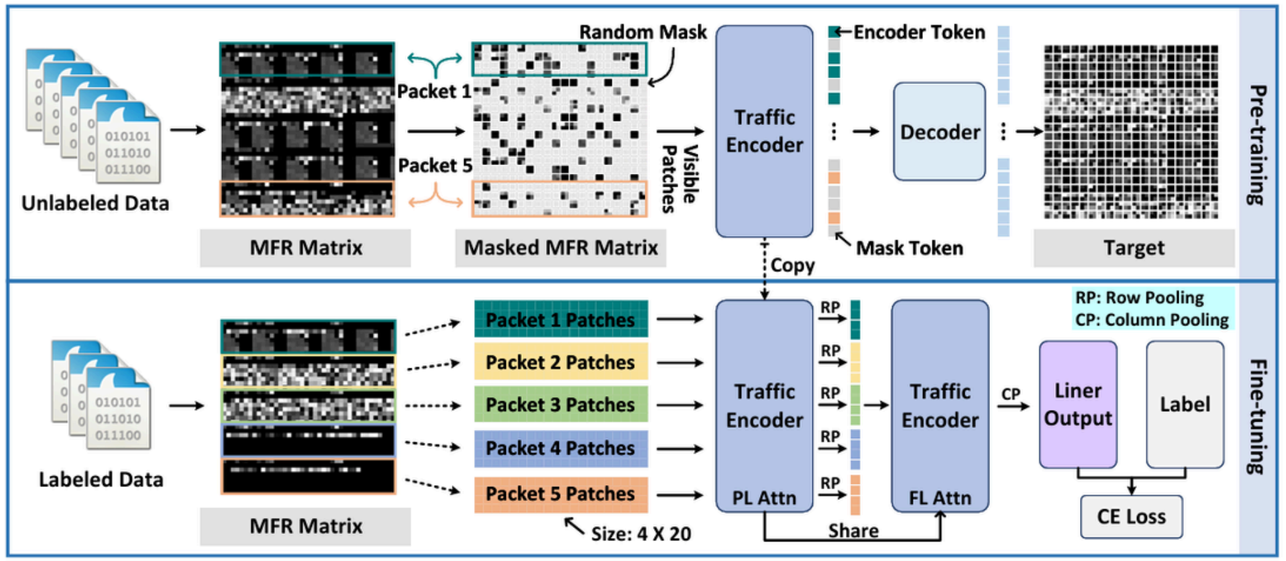
複雜度O(N^2)減少至O(N)

x_MFR = Pooling (x_c)

4.Loss function

Pre-training: $L_{rec} = \mathrm{MSE}(y_{rec}, y_{real})$

Fine-tuning: $L_{CE} = H(\hat{y}, y)$



實驗:
資料集:
CICIoT2022_MFR, ISCXTor2016_MFR,
ISCXVPN2016_MFR, USTC-TFC2016_MFR, Cross-Platform dataset (不參與訓練)

設定:
Pre-training batch size: 512, total step: 150000
Base learning rate: 1e-3, mask ratio: 90%
Fine-tuning optimizer: AdamW, epoch: 200,
Base learning rate: 2e-3, batch size: 64

比較對象:
FlowPrint, AppScanner, DF, Deeppacket, 2D-CNN, 3D-CNN
FS-Net, PERT, ET-BERT

實驗結果:
YaTC在各個資料集均取得最好的Acc和F1 score
YaTC在Cross-Platform資料集表現較好(泛用)
減少labeled data size時，YaTC效能降低最少
移除Packet-level Attention後性能大幅下降，
證明封包內的特徵提取的重要性
直接用raw bytes而非MFR後性能下降
驗證MFR轉換重要性

復現:
更新論文使用Python版本(3.8->3.11)，
補上epochs參數
新增擷取pcap檔案並轉成MFR圖片功能

	ISCXVPN_2016	ISCXTor2016	USTC-TFC2016	CICIoT2022_MFR
Acc	0.9545	0.9313	0.9562	0.9518
F1	0.9536	0.9316	0.9557	0.9517

復現環境:
CPU: Intel(R) Xeon(R) Gold 5218 CPU @ 2.30GHz
GPU: Quadro RTX 6000/RTX 8000
RAM: 396GB
OS: Debian(Linux 6.1.0-29-amd64)

Python 3.11.2
numpy 2.3.3
sympy 1.14.0
scipy 1.16.2
torch 2.8.0
torchvision 0.23.0
scikit-learn 1.7.2