Extending NetML's Flow2Features for Temporal and Rate-Based Video Streaming Inference

Group: Johnston Liu, Sriram Ananthakrishnan, Leon Luo

Project Summary

Existing network feature extraction libraries like NetML effectively transform packet captures into flow-level feature vectors, but their feature representations primarily capture static aggregates (e.g., total packets, mean IAT). For video streaming traffic, which exhibits bursty, segment-based patterns, these summaries fail to represent temporal variation that reflects playback quality. This project aims to extend NetML's flow2features module to incorporate segment-aware temporal and rate-based features, enabling machine learning models to better infer video quality from encrypted traffic. We will evaluate whether these new representations improve classification accuracy of streaming quality states compared to baseline NetML features.

Data

We will collect network traces from Netflix and YouTube video sessions using Wireshark and nPrint. Because these services encrypt video payloads, we will derive segment boundaries heuristically (e.g., time gaps >1 s between bursts) and label traces based on observable quality shifts (e.g., resolution changes, rebuffering events, bitrate plateaus). This approach provides a realistic testbed for encrypted traffic inference without requiring packet payload access.

Machine Learning

We will train models on both baseline and extended feature sets using: Random Forests and Gradient Boosted Trees (XGBoost, LightGBM) for tabular inference. Optionally, Temporal CNNs to test whether sequential segment data improves performance. Feature importance will be used to interpret which new features contribute most to model performance.

Evaluation

We will split data into train/validation/test sets across different streaming sessions. Evaluate using accuracy, F1-score, and ROC-AUC. Compare baseline NetML vs. extended NetML+ features using paired t-tests to test statistical significance. Perform ablation studies to assess each new feature's marginal contribution.

Proposed Feature Extensions

We will extend NetML's `flow2features()` by:

1. **Segment Download Rate** – Average bytes/sec for each inferred segment.
2. **Temporal Burst Metrics** – Standard deviation and coefficient of variation of segment durations.
3. **Retransmission and Loss Rates** – Derived from TCP flags to approximate congestion.
4. **RTT Statistics** – Mean and jitter computed from ACK timing.
5. **Inter-Arrival Time Moments** – Mean, variance, and skewness of IAT across flow windows.

6. **Configurable Windowing** – Allow time-based sampling in `SAMP_NUM`/`SAMP_SIZE` and consistent feature naming for downstream ML.
7. **Native Flow Filters** – Add IP-based filtering and modular feature selection.

Learning Objective

Empirical Insight: Determine which temporal or rate-based features best predict streaming quality.
Engineering Practice: Gain experience extending and contributing to open-source ML libraries (NetML).
Scientific Understanding: Explore how passive measurements can capture encrypted streaming dynamics through timing features.