Summary of Path-Embedding Signature Vector (PESV) Generation

Thesis Project Workflow

November 1, 2025

1 Objective

The primary objective of this project was to design and implement a novel, multi-modal feature vector, the **Path-Embedding Signature Vector (PESV)**, for the purpose of advanced VPN traffic classification. The final vector is a composite signature $\Sigma = (\alpha, \beta, \gamma)$, where each component captures a distinct characteristic of a network flow.

This document details the complete workflow, from initial data preprocessing to the final assembly of the complete PESV dataset.

2 Initial Data Preprocessing

The foundation of the project was built upon the ISCXVPN2016 dataset. The raw `.pcap` files were not suitable for direct analysis and required significant preprocessing, as documented in `P5_Dokumentasi.pdf`.

- **Flow Conversion:** Raw packet-level `.pcap` files were first converted into bidirectional flows (sessions) using the `SplitCap` tool. This step is critical for shifting the analysis from individual packets to entire network "conversations".
- **Data Cleaning:** A multi-stage filtering process was applied to the raw flows to remove noise and ensure data quality:
  1. Removal of OS-level protocol noise (e.g., DNS, NBSS, LLMNR).
  2. Verification of TCP flows to ensure they began with a complete 3-way handshake.
  3. Removal of specific UDP broadcast "Beacon˜" packets.
- **Final Dataset:** This cleaning and conversion process resulted in our starting dataset, located in the `new_flow/` directory, containing 5122 distinct flow `.pcap` files.

3 Component α: Learned Sequence Representation

3.1 Goal

To capture the underlying "grammar" and structural patterns of a flow by analyzing its sequence of packet sizes and directions.

3.2 Process

1. **Sequence Extraction:** For each of the 5122 flows, the first $N = 128$ packet sizes were extracted. Direction was preserved by encoding client-to-server packets as positive (`+size`) and server-to-client as negative (`-size`).
2. **Label Generation:** During the same extraction loop, a robust labeling function parsed each filename to extract three ground-truth labels: `application`, `category`, and `binary_type` (VPN/NonVPN).
3. **Autoencoder Training:**
   - The packet size sequences were normalized.
   - An LSTM-based autoencoder was built and trained in TensorFlow/Keras.
   - The model was trained to reconstruct its own input, forcing it to learn a compressed, 32-dimensional latent representation in its "bottleneck" layer.
4. **Feature Generation:** A new `encoder` model was created from the trained autoencoder's layers. All sequences were passed through this `encoder` to generate the final α feature vectors.

3.3 Outcome

`final_alpha_component_with_labels.csv`. This file contained **5084 rows**, as the script correctly skipped 38 files that were empty or did not contain IP packets.

4 Component β: Interarrival Time Distribution

4.1 Goal

To capture the "rhythm" of a flow by comparing the shape of its packet interarrival time (IAT) distribution against category-wide prototypes.

4.2 Process

1. **Challenge 1: Memory Crash:** The initial strategy involved loading all IATs from the training set into memory to build prototypes. This process crashed, exceeding 12GB of RAM.
2. **Solution: Histogram-based Prototypes:** A memory-efficient solution was implemented:
   - The training data was scanned once to find the global minimum and maximum IAT, establishing a fixed range for histogram bins.
   - On a second pass, normalized prototype histograms were built for each category by incrementally summing the histograms of their respective flows. This kept memory usage constant and low.
3. **Feature Generation:** For each of the 5122 flows, its own IAT histogram was created. The **Wasserstein distance** (Earth Mover's Distance) was then calculated between the flow's histogram and each of the category prototypes. This vector of distances became the β component.

4.3 Outcome

`final_beta_component_with_labels.csv` with **5122 rows**.

5 Component γ: Burstiness Profile Similarity

5.1 Goal

To capture the macro-level "conversational" dynamics of a flow by analyzing its burst statistics.

5.2 Process

1. **Burst Definition:** A burst was defined as a group of packets where the idle time between them was less than 1.0 second.
2. **Challenge 2: All-Zero Prototypes:** The first execution produced all-zero prototypes. A debug script revealed the cause:
   - **Root Cause:** A `TypeError` (`numpy.int64 to Decimal`). Scapy's `pkt.time` attribute returns a high-precision `Decimal` object, which is incompatible with NumPy's mathematical functions.
   - **Solution:** The script was fixed by immediately casting all timestamps to a standard `float(pkt.time)`, ensuring all subsequent math was compatible.
3. **Feature & Prototype Generation:**
   - The file-reading and feature extraction process was parallelized using `joblib.Parallel` to dramatically reduce runtime.
   - The script filtered all 5122 flows, identifying **4140 valid, multi-packet flows** suitable for burst analysis.
   - Prototypes were built by averaging four statistics (avg. packets/burst, avg. volume/burst, avg. duration/burst, avg. idle time) from the *training set* of these valid flows.
4. **Feature Generation:** The **Cosine Similarity** was calculated between each of the 4140 valid flow's burst vectors and the set of category prototypes. This vector of similarities became the γ component.

5.3 Outcome

`final_gamma_component_with_labels.csv` with **4140 rows** and a `filepath` column to serve as a unique key.

6 Final PESV Assembly

6.1 Goal

To combine the three feature sets (α, β, γ) into a single, unified dataset.

6.2 Process

1. **Challenge 3: Mismatched Row Counts:** The three CSVs had different lengths (5084, 5122, and 4140).
2. **Solution: Key-Based Merging:**

- The df_gamma (4140 rows) was used as the "base," as it contained the filepath key for all valid, analyzable flows.
- A robust script re-scanned the new_flow/ directory, simulating the skipping logic of the α and β scripts to generate their respective ordered filepath lists.
- These keys were added to the df_alpha (5084 rows) and df_beta (5122 rows) dataframes.
- An inner merge was performed, merging df_alpha and df_beta onto df_base_gamma using filepath as the common key.

6.3 Outcome

final_PESV_dataset.csv: A single, complete dataset with **4140 rows**. Each row contains the flow's labels and its complete α, β, and γ feature components, perfectly aligned and ready for model training.