Detection of Encrypted Traffic in Large Datasets

Aviv Turgeman 208007351
Yael Rosen 209498211
Adam Sin 322453689

November 2024

1 Abstract

The widespread use of encryption protocols like TLS and SSL has improved the security and privacy of network communications but has also complicated traffic classification. Traditional payload inspection methods no longer work, shifting the focus to metadata and advanced machine learning. This project tackles these challenges by developing a large, diverse encrypted traffic dataset and applying cutting-edge few-shot and zero-shot learning techniques. By addressing scenarios with limited labeled data, we aim to create adaptable and generalizable models to improve traffic analysis and management in encrypted environments.

2 Introduction

The increasing use of encryption technologies like TLS and SSL has greatly enhanced the security and privacy of network communications by concealing data content. This shift protects users from threats such as eavesdropping and data breaches, making encrypted traffic the dominant form of internet communication today.

However, encryption creates significant challenges for traffic classification, a key task for network management and security. Traditional methods like Deep Packet Inspection (DPI), which rely on analyzing packet content, are ineffective in encrypted environments. Without access to payload data, only metadata—such as packet size, timing, and flow patterns—can be analyzed, leaving critical gaps in identifying traffic, managing resources, detecting anomalies, and enforcing security policies.

Machine learning has shown potential in addressing these challenges. Few-shot learning enables models to work effectively with minimal labeled data, while zero-shot learning allows classification of traffic from previously unseen categories by leveraging semantic relationships. These methods are particularly useful in dynamic environments where new applications and services emerge regularly.

Our project main goals:

1. To create a comprehensive and diverse dataset of encrypted network traffic that reflects real-world scenarios and applications.

2. To apply few-shot and zero-shot learning techniques for classifying encrypted traffic with minimal labeled data.

By focusing on these objectives, this research seeks to develop scalable and adaptable solutions for encrypted traffic classification.

3 Related work

This section explores existing studies and highlights key research relevant to encrypted traffic classification, focusing on few-shot and zero-shot learning approaches. The following works were analyzed:

1. ***A Critical Study of Few-Shot Learning for Encrypted Traffic Classification***
   rboutaba.cs.uwaterloo.ca

   - **Abstract:**
     This study evaluates the effectiveness of few-shot learning methods like Matching Networks for encrypted traffic classification. It analyzes six real-world ISP datasets collected between 2019 and 2021.
   - **Conclusion:**
     Few-shot learning models show promising accuracy on specific datasets but struggle with generalization across different traffic environments, emphasizing the need for more robust methods.
   - **Datasets used:**
     ISCX-VPN2016, CIC-IDS2017.

2. ***Few-Shot Encrypted Traffic Classification: A Survey***
   ieeexplore.ieee.org

   - **Abstract:**
     This survey provides an overview of few-shot learning techniques for encrypted traffic classification, discussing methodologies, datasets, and research gaps.
   - **Conclusion:**
     The study highlights challenges like dataset diversity and model generalization, emphasizing the need for standardized datasets and approaches.
   - **Datasets used:**
     ISCX-VPN2016, CTU-13, UNSW-NB15.

3. ***Attribute-Based Zero-Shot Learning for Encrypted Traffic Classification***
ieeexplore.ieee.org

   - **Abstract:**
   Proposes an attribute-based zero-shot learning approach that uses a semantic space to classify unseen traffic types without additional labeled data.
   - **Conclusion:**
   The method effectively classifies unseen traffic, showcasing the adaptability of semantic attributes for encrypted traffic classification.
   - **Datasets used:**
   USTC-TFC2016, VNAT.

4. ***Encrypted Mobile Traffic Classification with a Few-Shot Incremental Learning***
researchgate.net

   - **Abstract:**
   Introduces a few-shot incremental learning model for encrypted mobile traffic that adapts to new patterns with minimal labeled data.
   - **Conclusion:**
   Demonstrates improved accuracy in dynamic environments, addressing the challenge of scarce labeled data for emerging applications.
   - **Datasets used:**
   ISCX-VPN2016, USTC-TFC2016.

5. ***Global-Aware Prototypical Network for Few-Shot Encrypted Traffic Classification***
ieeexplore.ieee.org

   - **Abstract:**
   Proposes a global-aware prototypical network for enhancing few-shot learning by incorporating global context into encrypted traffic classification.
   - **Conclusion:**
   Highlights improved robustness and accuracy for few-shot scenarios with global-aware mechanisms.
   - **Datasets used:**
   USTC-TFC2016, ISCX-VPN2016.

6. ***Many or Few Samples?***
arxiv.org

- **Abstract:**
  Analyzes the effect of training sample sizes on classification performance, focusing on trade-offs between large and few-shot datasets.
- **Conclusion:**
  Demonstrates that few-shot techniques can achieve competitive accuracy with fewer samples while balancing model complexity.
- **Datasets used:**
  CIC-IDS2017, UNSW-NB15, CTU-13.

7. ***High-Efficient and Few-Shot Adaptive Encrypted Traffic Classification***
   ieeexplore.ieee.org

   - **Abstract:**
     Introduces a few-shot adaptive classification model utilizing a DeepTree structure to optimize efficiency for real-time encrypted traffic analysis.
   - **Conclusion:**
     Highlights the model's high adaptability and efficiency, demonstrating its practicality for real-time applications.
   - **Datasets used:**
     VNAT, CIC-IDS2017, IoT-23.

4 Contribution

Our project focuses on advancing encrypted traffic classification by addressing the limitations of prior works and introducing innovative approaches. We aim to build a diverse and realistic encrypted traffic dataset while developing classification models that effectively use few-shot and zero-shot learning techniques.

4.1 How Our Work Differs from Previous Works

4.1.1 *Dataset Diversity and Realism:*

Current datasets like ISCX-VPN2016 and CIC-IDS2017 are limited in scope. We aim to integrate broader applications and real-world scenarios to create a richer, more representative dataset.

4.1.2 *Few-Shot and Zero-Shot Learning:*

Existing models face challenges in generalizing across datasets. Our focus is on building adaptive models that classify unseen traffic types with minimal labeled data, ensuring they perform well in diverse environments.

4.1.3 *Feature Extraction and Efficiency:*

While feature extraction is a focus in works like "DeepTLS," we emphasize balancing efficiency with accuracy. Our approach will incorporate optimized time-series analysis and adapt techniques like FlowPic for few-shot learning.

4.1.4 *Integration of Innovations:*

Unlike previous efforts that isolate datasets or models, we combine advanced dataset creation, machine learning techniques, and practical real-world application into one cohesive framework.

4.2 Research Goals of the Project

4.2.1 *Dataset Creation:*

Build a comprehensive dataset that reflects diverse traffic types and real-world conditions, supporting both PCAP and metadata-based analyses.

4.2.2 *Innovative Classification Models:*

Develop few-shot and zero-shot models that classify traffic with minimal data and handle unseen traffic types across scenarios like VPN, Tor, and IoT traffic.

4.2.3 *Scalability and Application:*

Create lightweight, real-time models suitable for IoT and constrained environments, while exploring adaptive systems that dynamically learn new patterns.

4.2.4 *Benchmarking and Standardization:*

Provide benchmarks for classification techniques and propose best practices for dataset creation and model evaluation to improve research consistency.

5 Background

Encrypted network traffic classification has become increasingly critical as encryption technologies grow, challenging traditional analysis methods. This section explores the core challenges and methodologies relevant to the field, emphasizing few-shot and zero-shot learning.

5.1 Encrypted Traffic: Challenges and Opportunities

Encrypted traffic hides payloads, leaving only metadata like packet size and flow patterns available for analysis. This enhances privacy and security but complicates tasks like classification and anomaly detection. New approaches are essential for analyzing encrypted traffic without compromising user privacy.

5.2 Traditional Traffic Classification Methods

Traditional methods such as Deep Packet Inspection (DPI) depend on analyzing packet content, which is no longer viable due to encryption. Flow-based methods, using features like packet size and timing, offer alternatives but struggle to generalize across dynamic traffic scenarios, creating a need for advanced machine learning.

5.3 Few-Shot and Zero-Shot Learning

Few-shot and zero-shot learning are recent advancements that address limited data and unseen categories.

- **Few-Shot Learning:** Enables learning new tasks with minimal labeled examples. Methods like Prototypical Networks and Matching Networks are effective in extracting representative embeddings and comparing new samples.
- **Zero-Shot Learning:** Classifies unseen classes using semantic relationships between known and unknown categories. This is critical for identifying emerging applications in encrypted traffic.

5.4 Feature Extraction for Encrypted Traffic

Effective feature extraction is key to traffic classification in encrypted environments. Common features include:

- **Statistical Features:** Packet size distribution and flow byte count.
- **Temporal Features:** Inter-arrival times and flow durations.
- **Protocol-Specific Features:** Metadata from TLS handshakes, such as record sizes.
- **Flow-Level Aggregation:** Summarized flow behaviors.

These features enable differentiation between traffic types without relying on payload analysis.