Network Packet Classification Using CIC Darknet, Tor, and VPN Datasets

Group 1

November 8, 2024

1 Introduction

1.1 Background

With the rise in encrypted and anonymized network traffic, effective classification of network packets has become an essential component of cybersecurity. Network activities through anonymizing channels, such as Darknet, Tor, and VPN, are often shielded from traditional detection mechanisms. This limitation poses significant challenges for network security, as cybercriminals leverage these channels to obscure malicious activities. Accurate classification of network packets can support threat intelligence and network defense operations by enabling detection of potentially suspicious or unauthorized traffic patterns. This study investigates the use of Convolutional Neural Networks (CNNs) as a modern approach for categorizing encrypted network traffic using datasets sourced from the Canadian Institute for Cybersecurity (CIC). Our focus is on examining encrypted traffic across three types of channels: Darknet, Tor, and VPN, each with unique characteristics and behavioral patterns.

1.2 Objective

The primary objective of this project is to develop a generalized model that can classify network packets across multiple types of encrypted traffic with high accuracy. We aim to leverage CNN architectures due to their effectiveness in extracting spatial hierarchies in data and capturing complex patterns within network traffic. By training on diverse datasets (Darknet, Tor, and VPN), we intend to create a robust classification model that can accurately identify traffic type and support cybersecurity operations by providing an additional layer of network visibility and aiding in the detection of suspicious encrypted traffic. This generalization of CNNs across various datasets aligns with our goal of developing a reliable, adaptable approach for monitoring encrypted network traffic in real-world cybersecurity contexts.

2 Datasets

In this study, we utilized three datasets provided by the Canadian Institute for Cybersecurity (CIC). These datasets contain network traffic from different anonymized sources—Darknet, Tor, and VPN—each reflecting distinct types of encrypted or anonymized internet traffic.

- **Darknet2020 Dataset**: Darknet is the unused address space of the internet which is not speculated to interact with other computers in the world. Any communication from the dark space is considered sceptical owing to its passive listening nature which accepts incoming packets, but outgoing packets are not supported. Due to the absence of legitimate hosts in the darknet, any traffic is contemplated to be unsought and is characteristically treated as probe, backscatter or misconfiguration.
- **Tor Dataset**: This dataset captures traffic generated within the Tor network, a prominent tool for online anonymity. The Tor network is commonly used for privacy-focused activities and provides encrypted pathways to prevent tracking of IP addresses and user activity.
- **VPN Dataset**: The VPN dataset consists of network traffic routed through Virtual Private Networks (VPNs), which are commonly employed for both legitimate privacy purposes and to bypass geographic restrictions. VPN traffic is encrypted and often used to obscure the user's real IP address, making it resemble Tor and Darknet traffic in terms of anonymity.

The diversity and specificity of these datasets make them suitable for building a CNN model capable of generalizing across various encrypted traffic types. By leveraging the unique characteristics found in each dataset, our approach aims to establish a model that can accurately classify packets as either Non-Darknet, Tor, or VPN,

3 Methodology

Our approach to this project involved a systematic exploration and application of models and techniques that have been previously successful in similar network classification tasks. The methodology unfolded in three phases: Model Selection, Model Implementation, and Feature Extraction through PCAP-to-CSV Conversion.

3.1 Phase I: Model Selection

We started by examining research papers associated with each dataset to understand the methods previously used for classifying network traffic. By leveraging these insights, we shortlisted a range of potential models before choosing the most promising approach.

3.1.1 Tor and NonTor Classification

In earlier research, classification between Tor and non-Tor traffic was achieved using feature selection techniques like CfsSubsetEval with BestFirst (SE+BF) and InfoGain with Ranker (IG+RK). These methods were used in combination with algorithms like ZeroR, C4.5 Decision Trees, and K-Nearest Neighbors (KNN) to determine their efficacy in distinguishing Tor from non-Tor traffic. Among these, the InfoGain + Ranker approach demonstrated superior performance in selecting relevant features, while the C4.5 model, coupled with IG+RK, yielded the highest accuracy. This approach also successfully classified application-specific traffic types, such as Browsing, Audio, and VOIP, with Random Forest and IG+RK providing optimal results. These insights highlighted the importance of effective feature selection when working with anonymized traffic.

3.1.2 VPN and Non-VPN Classification

For VPN classification, researchers used time-related features with C4.5 Decision Trees and KNN. Both algorithms showed high accuracy, with C4.5 performing slightly better in identifying VPN versus non-VPN traffic. This model selection reinforced that time-based features are critical in distinguishing VPN traffic patterns, which are inherently different from non-VPN traffic due to their encryption protocols.

3.1.3 Darknet Classification

A novel technique called DeepImage was introduced for Darknet traffic classification, where feature selection was used to create grayscale images from the data. These images were then fed into a two-dimensional CNN, achieving an impressive 86% classification accuracy for Darknet traffic. DeepImage's transformation of network features into visual representations enabled the CNN to leverage spatial patterns within the data, making it highly effective for identifying complex traffic behaviors unique to Darknet traffic. This robust performance led us to adopt a CNN-based model, specifically the DeepImage approach, as our foundational model due to its adaptability and high classification accuracy.

3.1.4 Final Model Selection

Based on these findings, the CNN-based DeepImage model was chosen for its robustness and adaptability across network traffic types. Furthermore, the Darknet2020 dataset was ideal for training, as it contained the necessary labels and traffic diversity. This selection laid a strong foundation for developing a model that could generalize well across multiple traffic types.

3.2 Phase II: Model Implementation

After selecting the DeepImage CNN model, we implemented and tested it on the Darknet2020 dataset, observing encouraging results in initial trials. This motivated us to apply the model to the Tor and VPN datasets as well. However, the CSV files for the Tor and VPN datasets lacked certain key features available in the Darknet2020 dataset. Despite this limitation, the CNN model performed adequately on the Tor dataset, yielding reasonable classification accuracy even with partial feature sets.

Metric	Darknet Dataset	Tor Dataset	VPN Dataset
Accuracy	95%	87%	48%
Precision	90.3%	82%	54%
Recall	88.7%	69.5%	50%
F1 Score	89.3%	75.2%	51.9%

Table 1: Initial Model Performance Metrics Across Datasets

These initial results confirmed that the CNN model could classify across datasets with fair accuracy. In order to fairly evaluate the VPN and Tor datasets and to maximize model performance, we recognized the need to extract all relevant features for Tor and VPN datasets, ensuring consistency in the feature set across all datasets.

3.3 Phase III: PCAP to CSV Conversion

To address the feature inconsistency, we utilized CICFlowMeter, an open-source tool for extracting detailed network flow features from packet capture (PCAP) files. By converting PCAP files to CSV, we generated labeled datasets with consistent feature sets, thus providing the model with comprehensive data inputs for both training and testing. This feature augmentation notably improved classification accuracy, as shown in the subsequent results. The extraction process allowed us to standardize features across Darknet, Tor, and VPN datasets, enhancing the model's ability to recognize and classify packet patterns with greater accuracy.

4 Results

Our results demonstrate that the CNN-based DeepImage model, when combined with consistent feature sets across the Darknet, Tor, and VPN datasets, yields strong classification performance. This section outlines the specific outcomes observed during our evaluation, focusing on accuracy, precision, recall, and F1 scores across datasets.

4.1 Performance Metrics

To assess the effectiveness of our model, we evaluated it on several key metrics: Accuracy, Precision, Recall, and F1 Score.

Metric	Non-Darknet	Tor	VPN
Accuracy	97%	97%	97%
Precision	99%	97%	80%
Recall	98%	92%	87%
F1 Score	98%	94%	83%

Table 2: Final Model Performance Metrics Across Datasets

Note that the DeepImage based CNN has a robust classification criterion which allows it to classify the data well across all features.

4.2 Impact of Feature Consistency

Upon converting the Tor and VPN datasets from PCAP to CSV files using CICFlowMeter, we observed a notable improvement in model accuracy. This process increased feature consistency across datasets, allowing the CNN to detect subtle network traffic patterns that were previously overlooked due to missing data. Table 3 shows the accuracy improvement across each label after feature extraction.

Dataset	After Feature Extraction	Before Feature Extraction
Non-Darknet	98%	84.67%
Tor	94%	71%
VPN	83%	43%

Table 3: F1-Score Improvements Due to Feature Consistency Across Datasets

These results affirm the importance of feature richness in training effective classification models, particularly for anonymized traffic.

4.3 Limitations and Considerations

While the CNN model demonstrates high accuracy, there are limitations worth noting:

- **Feature Variability:** Despite feature extraction, minor inconsistencies in packet patterns can limit model generalizability across diverse network environments.
- **Real-Time Application:** For practical, real-time classification, further optimization would be required to minimize processing delays.

- **VPN Recognition:** The model performs slightly better for Tor or Non-Darknet than for VPN entries. Further time-based feature analysis using techniques such as Random Forest (as used in the research paper) may be used to enhance VPN label recognition.

These limitations suggest potential areas for future improvements, such as the use of larger datasets with a more balanced distribution of traffic types and feature engineering to enhance model robustness across varied network scenarios.

5 Future Improvements

In order to enhance the performance and applicability of the model, several future improvements can be explored:

- **Application-Based Classification**: Currently, the model classifies traffic into broad categories such as Tor, VPN, and Darknet. However, a more granular classification based on specific application types (e.g., Browsing, Audio, Chat, Mail, Peer-to-Peer (P2P), File Transfer (FTP), VOIP, and Video) could improve the model's versatility and provide more detailed insights into network traffic.
- **Hybrid Models**: Another potential direction is the integration of Convolutional Neural Networks (CNN) with Recurrent Neural Networks (RNN), specifically Long Short-Term Memory (LSTM) networks, to create a hybrid model. While CNNs are effective at extracting spatial features from network packets, RNNs (particularly LSTMs) excel in capturing temporal dependencies in sequences of data. Network traffic often exhibits time-dependent patterns (as seen in the VPN research paper), such as periodicity in data flow or burst traffic, which might be better captured by an RNN.
- **Real-Time Adaptation**: To make the model more suitable for practical deployment in real-world environments, it can be modified for real-time classification. This involves optimizing the model for lower latency and faster processing, ensuring that it can handle high volumes of network traffic without delays. Real-time adaptation would be particularly beneficial in applications such as intrusion detection systems (IDS).

6 References

1. Darknet2020 Dataset
2. Tor Dataset
3. VPN Dataset
4. GitHub - Onion-Tunnel-CS658