Analysis of the ISCX VPN-nonVPN Dataset 2016 for Encrypted Network Traffic Classification

Felipe Peter
Tsinghua University
felipe.peter@tum.de

Abstract

Already half of today's internet traffic is encrypted using protocols like SSL/TLS. This prevents classic deep packet inspection approaches from analyzing packet payloads. Recently, researchers published a deep learning approach, claiming that their trained model is capable of finding patterns in encrypted network traffic payloads and classifying applications based on these patterns. This work shows that the claim is unlikely to be true, because the utilized dataset exposes features that allow for highly accurate classification without incoporating any payload data.

1. Introduction

Packet classification is an essential function to enable functions of the internet, like firewall, access control, traffic monitoring and billing, policy-based routing, and traffic rate limiting [7, 15].

Packet classification approaches can be categorized according to the "depth" of packet inspection, which increased over the past decades [13]. The shallowest way to control IP packets is by looking at the IP header and blocking or forwarding accordingly. As the lookup of the header information takes a comparatively long time, stateful packet inspection was invented. A stateful inspection caches so called "flows" to speed up rule lookup for subsequent packets of the same flow. A typical choice of header fields is the TCP/UDP 5-tuple consisting of source IP address, source port, destination IP address, destination port, and transport protocol type. The field values are usually IP address prefixes, exact values or wildcards for protocols, and ranges for port numbers [15].

This port-based classification has become less accurate over the years because applications started to use random ports or well-known ports assigend to other applications, in order to disguise their traffic and circumvent packet classification. Furthermore, the exhaustion of the IPv4 address space has lead to increased usage of network address translation solutions to offer web access to several servers or clients by sharing IP addresses on different ports [6].

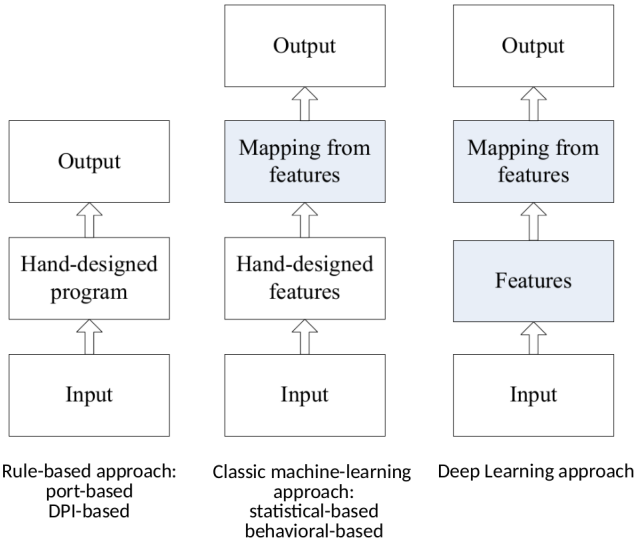

Figure 1. Taxonomy of traffic classification approaches. Deep learning is the only approach that does not need hand-crafted features, but learns them as part of the end-to-end learning procedure. (Depiction adapted from [18].)

Application proxies use medium depth packet inspection to improve security. An application proxy works as an intermediary between clients and servers and can enforce rules by inspecting application layer protocols like HTTP and FTP [11].

Deep packet inspection (DPI) was invented to overcome the limitations of shallow and medium packet inspection. It takes the entire IP packet as input and uses different analyses to make forwarding decisions. The payload can be compared against databases of predefined strings or signatures that can be associated with attacks, e.g. with malicious software. Additionally, the static pattern matching can be supplemented by analyzing statistics like packet lengths or historical timeseries information about a connection [13]. Deep packet inspection is computationally more demanding than header inspection because the headers are, in contrast to the payload, restricted in position, length, and content by

Traffic Type	Content
Email	Email、Gmail (SMPT, POP3,IMAP)
VPN-Email	
Chat	ICQ, AIM, Skype, Facebook, Hangouts
VPN-Chat	
Streaming	Vimeo, Youtube, Netflix, Spotify
VPN-Streaming	
File transfer	Skype, FTPS, SFTP
VPN-File transfer	
VoIP	Facebook, Skype, Hangouts, Voipbuster
VPN-VoIP	
P2P	uTorrent, Bittorrent
VPN-P2P	

Figure 2. Applications and traffic types of the ISCX VPN-nonVPN dataset 2016 according to Wang et al. [17].

protocol standards.

In recent years, many websites and services on the web have transitioned from HTTP to HTTPS, using the SSL/TLS encryption protocol. According to the Cisco 2018 Annual Cybersecurity Report [4], 50% of the global internet traffic has been encrypted in October 2017. That is a 12-point increase from November 2016. The encryption of web traffic poses serious problems for DPI, as the payload can no longer be examined in its unencrypted form. After big companies like Facebook and Google moved their services to HTTPS in 2011 and 2012 respectively, DPI in its original form was considered to be rendered useless [5].

As a result, the usage of statistical-based and behavioral-based approaches increased [18]. Statistical-based approaches leverage the diversity of web traffic (e.g., short packet bursts in Voice over IP applications in contrast to long, steady flows of file downloads). Examined features can be packet length, number of packets in a given time, or used protocol. Behavioral-based approaches on the other hand, examine an endpoint in a network and analyze for example how many hosts are contacted, with which protocol and on how many different ports [3]. These approaches originally used hand-crafted features and "classic" machine-learning techniques, like k-nearest neighbor or random forests [14]. The advent of real-life applicability of neural networks since 2012 [9] promoted research on using deep neural networks (deep learning) to solve the traffic classification problem with an end-to-end approach. Deep learning models usually do not use hand-crafted features but directly learn the underlying statistics of the input data. This enables them to learn very complex functions which makes them outperform classic machine-learning techniques. A schematic comparison of the traffic classification approaches can be seen in Fig 1.

2. Related Work

The research on deep learning for encrypted traffic classification was mainly conducted since 2017. Several approaches were proposed, some based on statistical and time-series features [8, 12] and some based on header and payload data [16, 2]. Furthermore, approaches can be categorized into those looking at a flow of several packets or only investigating a single packet at a time. Rezaei and Liu [14] wrote a survey on the comparison of these approaches.

Application	CNN			SAE		
	Rc	Pr	F_1	Rc	Pr	F_1
AIM chat	0.76	0.87	0.81	0.64	0.76	0.70
Email	0.82	0.97	0.89	0.99	0.94	0.97
Facebook	0.95	0.96	0.96	0.95	0.94	0.95
FTPS	1.00	1.00	1.00	0.77	0.97	0.86
Gmail	0.95	0.97	0.96	0.94	0.93	0.94
Hangouts	0.98	0.96	0.97	0.99	0.94	0.97
ICQ	0.80	0.72	0.76	0.69	0.69	0.69
Netflix	1.00	1.00	1.00	1.00	0.98	0.99
SCP	0.99	0.97	0.98	1.00	1.00	1.00
SFTP	1.00	1.00	1.00	0.96	0.70	0.81
Skype	0.99	0.94	0.97	0.93	0.95	0.94
Spotify	0.98	0.98	0.98	0.98	0.98	0.98
Torrent	1.00	1.00	1.00	0.99	0.99	0.99
Tor	1.00	1.00	1.00	1.00	1.00	1.00
VoipBuster	1.00	0.99	0.99	0.99	0.99	0.99
Vimeo	0.99	0.99	0.99	0.98	0.99	0.98
YouTube	0.99	0.99	0.99	0.98	0.99	0.99
Wtd. Average	**0.98**	**0.98**	**0.98**	**0.96**	**0.95**	**0.95**

Table 1. Results of the Deep Packet approach. The convolutional neural network architecture outperforms the stacked auto-encoder. [10]

This work will focus on two papers proposed in 2017 and produced in parallel by Wang et al. [17] and Lotfollahi et al. [10]. Both approaches use one-dimensional CNN (Lotfollahi et al. also test a stacked auto-encoder) on single packets, including header and payload. Both papers use the ISCX VPN-nonVPN dataset 2016 [1] for training and testing their models. The ISCX VPN-nonVPN dataset 2016 consists of around 30GB of recorded traffic. It can be used for traffic identification or application classification. For this reason the used applications have been grouped into traffic groups. Fig. 2 shows the applications and traffic types as used in the paper by Wang et al. [17]. Lotfollahi et al. [10] merge uTorrent and Bittorrent into one Torrent application.

Wang et al. published their paper first, reporting high accuracy for the traffic identification task. Lotfollahi et al. doubt these results, because Wang et al.'s approach incorporates the source and destination IP address of packets. These are unique for the applications in the dataset. To avoid this behaviour, Lotfollahi et al. mask the IP addresses of all packets. They report high accuracy for both the traffic identification and application classification task, as shown in Table 1. The CNN approach outperforms the approach using a stacked auto-encoder. The authors claim that their network, that they name Deep Packet, learned to find discriminative patterns in the payloads based on the pseudo-random encryption schemes of each application. The following Chapter will further investigate whether that claim holds true.

Application	Number of flows	Number of packets	Unique flows	Unambiguous packets
AIM chat	424	4248	0.12	0.2
Email	2775	38464	0.28	0.41
Facebook	21726	2586323	0.15	0.89
FTPS	1087	5911479	0.89	1.0
Gmail	250	9931	0.74	0.94
Hangouts	20897	3792385	0.14	0.93
ICQ	455	3724	0.12	0.13
Netflix	161	160430	0.92	1.0
SCP	5769	415955	0.26	0.93
SFTP	173	782127	0.68	1.0
Skype	24843	3663605	0.24	0.91
Spotify	167	22129	0.95	0.98
Torrent	1468	72337	0.87	0.98
Tor	42	210605	0.88	1.0
VoipBuster	3090	841590	0.41	0.98
Vimeo	424	92683	0.95	1.0
Youtube	493	137941	0.97	0.97
Total/ Average	84244	18745956	0.23	0.95

Table 2. Results of the dataset analysis based on flows consisting of the triplet source port, destination port, and used protocol. While the fraction of unique flows is relatively low, 95% of the packets can be uniquely associated with one application. This analysis incorporates only non-VPN traffic.

3. Dataset Analysis

This work focuses on the same packets that were used in the Deep Packet paper. That means that only TCP and UDP packets with an application related payload were considered. Therefore, DNS packets and SYN, ACK, FIN packets were discarded. Lotfollahi et al. have not published code for the data preprocessing stage, but based on the number of packets they mention for every application it is reasonable to assume that they only used non-VPN traffic for the application classification task.

To examine the flows of the dataset, every packet header is inspected. But instead of looking at the 5-tuple consisting of source IP, destination IP, source port, destination port, and protocol, only the source port, destination port, and protocol are used to define a flow. This accounts for the IP masking step in the Deep Packet paper. As this eliminates the uniqueness of a flow in the dataset by removing the application specific IP address, flows should not be associable with applications anymore. Table 2 shows the results of this analysis. Only 23% of the non-VPN flows in the dataset are unique and could therefore be classified as belonging to one application. All other flows are used by at least two applications. Only a few applications, like Youtube or Spotify, could be identified with high accuracy based on a given flow. This suggests that the Deep Packet approach should not be able to classify applications with high accuracy only by looking at the header information. However, taking into account the number of packets per flow leads to a different conclusion. The unique flows include 95% of all packets in the dataset. That means that indeed, one can get a very high classification accuracy by simply mapping flows to applications. If a packet belongs to a non-unique flow, guessing the application still has a relatively high chance of being correct, therefore increasing the accuracy beyond 95%.

Table 2 shows only non-VPN traffic. If the whole dataset is taken into account, the fraction of unique flows is higher, but the number of packets that can be associated with a specific app decreases to 92% (see Table 3). Again, guessing the application for ambiguous packets increases the classification accuracy. However, the whole dataset is only used for traffic type identification in the Deep Packet paper.

4. Discussion

Comparing the results of Tables 1 (CNN results) and 2 shows a correlation between the predictive performance of the Deep Packet model and the amount of unambiguous packets that can be associated with each application. The applications with low F1-scores are also the applications that have a very low percentage of unambiguous packets. Namely AIM chat and ICQ have the lowest F1-scores and are also the applications with the lowest percentage of unambiguous packets, followed by Email.

It is very unlikely that the Deep Packet approach actually learned how to find patterns in the encrypted data. It is more likey that the model learned to memorize the mapping from flows to applications.

Using the whole dataset (VPN and nonVPN) reduces the accuracy for application classification (see Table 3) but as applications are grouped into traffic types and grouped applications probably share flows, the traffic type identification performed with the Deep Packet approach most likely also just memorizes the flow-to-traffic-group mapping.

Application	Number of flows	Number of packets	Unique flows	Unambiguous packets
AIM chat	458	5601	0.13	0.37
Email	3057	51098	0.33	0.55
Facebook	22253	2601864	0.17	0.88
FTPS	1291	6038433	0.9	1.0
Gmail	250	9931	0.74	0.94
Hangouts	21613	4689219	0.16	0.94
ICQ	495	8096	0.15	0.5
Netflix	286	732836	0.86	0.23
SCP	5769	415955	0.26	0.93
SFTP	196	855456	0.72	1.0
Skype	26473	4466639	0.28	0.92
Spotify	299	97470	0.93	0.25
Torrent	2097	341433	0.9	0.95
Tor	42	210605	0.86	1.0
VoipBuster	3823	1563872	0.36	0.99
Vimeo	551	366694	0.89	0.86
Youtube	803	268310	0.95	0.69
Total/ Average	89756	22723512	0.27	0.92

Table 3. Results of the dataset analysis based on flows consisting of the triplet source port, destination port, and used protocol. While the fraction of unique flows is relatively low, 92% of the packets can be uniquely associated with one application. This analysis incorporates both VPN and non-VPN traffic.

Overall the situation in current research regarding encrypted traffic classification is methodically unsatisfying. Authors do not publish code, use non-public datasets, or do not explain data preprocessing steps in detail. This makes it hard to evaluate their results. Nonetheless, usage of statistical and timeseries data seems to be the state-of-the-art solution for encrypted traffic classification.

5. Conclusion

This work has shown that an approach using the ISCX VPN-nonVPN dataset 2016 for packet-level encrypted traffic classification can not incorporate packet header information, as it allows to directly map a packet to a specific application with high accuracy. Considering only non-VPN traffic, 95% of all packets in the dataset can be associated with an application. The remaining packets can still be classified with high probability by guessing based on the applications that use this flow. This result suggests that it is very likely that the Deep Packet approach proposed in [10] learns to memorize the mapping from flows to applications instead of finding application-specific encryption patterns.

While the ISCX VPN-nonVPN dataset might be suitable for statistical and timeseries approaches, future work should investigate real-world traffic in order to generate better datasets for packet-level classification.

References

[1] 2nd International Conference on Information Systems Security and Privacy(ICISSP 2016). *Characterization of Encrypted and VPN Traffic Using Time-Related Features*, 2016.

[2] G. Aceto, A. Montieri, D. Ciuonzo, and A. Pescapè. Mobile encrypted traffic classification using deep learning, May 2018.

[3] E. Biersack, C. Callegari, and M. Matijasevic. *From Measurement, Classification, and Anomaly Detection to Quality of Experience*, volume 7754 of *Computer Communication Networks and Telecommunications*. Springer-Verlag Berlin Heidelberg, 1 edition, 2013.

[4] Cisco. Cisco 2018 annual cybersecurity report. Technical report, Cisco, 2018.

[5] M. Coward. Encryption: will it be the death of dpi? http://telecoms.com/39718/encryption-will-it-be-the-death-of-dpi/, February 2012.

[6] A. Dainotti et al. Issues and future directions in traffic classification. *IEEE Network*, January 2012.

[7] P. Gupta and N. McKeown. Packet classification on multiple fields, 1999.

[8] J. Höchst, L. Baumgärtner, M. Hollick, and B. Freisleben. Unsupervised traffic flow classification using a neural autoencoder. 10 2017.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[10] M. Lotfollahi, R. Shirali, M. Jafari Siavoshani, and M. Saberian. Deep packet: A novel approach for encrypted traffic classification using deep learning. *ISeCure*, 09 2017.

[11] A. Luotonen and K. Altis. World-wide web proxies. http://courses.cs.vt.edu/ cs4244/spring.09/documents/Proxies.pdf, April 1994.

[12] E. Mahdavi, A. Fanian, and H. Hassannejad. Encrypted traffic classification using statistical features. *ISeCure*, 10(1):29–43, January 2018.

[13] T. Porter. The perils of deep packet inspection. https://www.symantec.com/connect/articles/perils-deep-packet-inspection, January 2005.

[14] S. Rezaei and X. Liu. Deep learning for encrypted traffic classification: An overview. *IEEE Communications Magazine Proposals*, October 2018.

[15] D. E. Taylor. Survey & taxonomy of packet classification techniques. Technical report, ACM COMPUTING SURVEYS, 2004.

[16] W. Wang, Y. Sheng, J. Wang, X. Zeng, X. Ye, Y. Huang, and M. Zhu. Hast-ids: Learning hierarchical spatial-temporal features using deep neural networks to improve intrusion detection. *IEEE Access*, December 2017.

[17] W. Wang, M. Zhu, J. Wang, X. Zeng, and Z. Yang. End-to-end encrypted traffic classification with one-dimensional convolution neural networks, July 2017.

[18] W. Wang, M. Zhu, X. Zeng, X. Ye, and Y. Sheng. Malware traffic classification using convolutional neural network for representation learning. In *2017 International Conference on Information Networking (ICOIN)*, pages 712–717, Jan 2017.