Traffic Classification using Deep Learning: Being Highly Accurate is Not Enough

Kang-Hee Lee
Sangmyung University
lkh7054@gmail.com

Seung-Hun Lee
Sangmyung University
mr.leesh90@gmail.com

Hyun-Chul Kim*
Sangmyung University
hyunchulk@gmail.com

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Networks** → **Network measurement**;

KEYWORDS

Deep Learning, Traffic Classification, eXplainable AI

1 INTRODUCTION

As Deep Learning (DL) algorithms have rapidly become a methodology of choice in various domains, they have recently entered also the field of the Internet traffic classification, successfully demonstrating impressive results. Most of the research work up to this point has focused on improving the accuracy of classification systems, yet there has been little attempt to provide (i) systematic comparison of the various DL algorithms used and (ii) analysis on where the higher accuracy come from, particularly when comparing with the traditional machine learning algorithms like C4.5. To fill this gap, we conduct experiments with four DL algorithms proposed for traffic classification, including CNN, LSTM, Stacked Auto-Encoder (SAE), and Hierarchical Attention Networks (HAN). Further, we propose to leverage and visualize hierarchical attention layers to highlight which parts of the traffic packet traces were most informative for accurate classification, which provides hints about why (and how) DL algorithms achieve the state-of-the-art level high accuracy. We view this paper as the first step towards answering the aforementioned "why" question, which is critical in understanding the real benefit and contribution of deep learning to the field of the Internet traffic classification, and advancing its state-of-the-art.

2 METHODOLOGY

CNN is known to have a powerful ability to extract and learn the spatial features from given data on a pixel-by-pixel basis, thus often used in the area of computer vision for image classification or object detection. LSTM is good at processing sequential data just like traffic traces. Auto-Encoder is an unsupervised learning method generally used for automatic feature extraction. SAE is a structure in which several auto-encoders are overlapped. HAN [3], originally proposed for document classification, has a hierarchical structure that builds two-level representations, at the word-level and sentence-level (i.e., bytes-level and packet-level, respectively, when applied for traffic classification). Moreover, it has attention mechanisms at both levels to be able to pay more attention to important words and sentences when constructing the document representation, which calculates the attention weights for each sentence and word during the learning process. Words or sentences with higher attention weight mean that they played a discriminative role in the classification process, thus we leverage this mechanism to find out what packets (in a given traffic flow) and bytes (in a packet) were informative for the algorithm's accurate classification decisions, by visualizing these attention layers.

Table 1: Application categories

WIDE	USTC	CTU
NTP	Facetime	Cridex
Web	Gmail	Geodo
SSH/SSL	MySQL	Htbot
BitTorrent	BitTorrent	Miuref
DNS	Outlook	Neris
MAIL	SMB	Nsis-ay
	FTP	Shifu
	Skype	Tinba
	Weibo	Virut
	WorldOfWarcraft	Zeus

We implemented and tested all the four DL algorithms on three traces with payload collected from different geographic locations, Japan (WIDE [2]), China (USTC [1]), and Czech (CTU [4]). We use the definition of a flow based on its 5-tuple (source IP address, destination IP address, protocol, source port, destination port). IP address and MAC address fields are all replaced with zeroes, so as not to make the information get automatically extracted and learned as a distinguishing pattern by DL algorithms. Table 1 summarizes application categories contained in each dataset. We randomly sample 5,000 flows for each application. We use 80% of the data for training, 10% for validation, and the remaining 10% for test.

3 PRELIMINARY RESULTS

Table 3 shows that the overall accuracy (i.e., the ratio of the sum of all True Positives to the sum of all the True Positives and False Positives for all classes) of all the tested DL algorithms, which ranges from 96.3% to 100%. HAN achieves the highest accuracy on every trace, with >=98.5%.

Table 2 shows an example visualization of the HAN's attention weight analysis results, with which we were able to locate which byte sequences and packets contributed the most in classifying the given flows. For the Web traffic flow, the fourth packet is carrying

*Corresponding Author

SIGCOMM '20 Demos and Posters, August 10–14, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-8048-5/20/08…$15.00
https://doi.org/10.1145/3405837.3411369

Table 2: HAN Attention Visualization

Application	Packet No.	Packet attention	Bytes attention	Information
Web	all	0.04	... 0x22 0x9e 0x00 0x50 0x08 0xa1 ...	Source Port: 80
	4	0.94	... 0x48 0x54 0x54 0x50 0x2f 0x31 ...	Payload: HTTP/1.1
BitTorrent	2	0.83	... 0x70 0x72 0x6f 0x74 0x6f 0x63 0x6f 0x6c ...	Payload: protocol
BitTorrent	1	0.87	... 0x18 0x98 0xb8 0xbe 0xca ...	Window size value: 39096

Table 3: Overall accuracy of the DL models

Traces	SAE	CNN	LSTM	HAN
WIDE	97.2%	98.5%	99.4%	100%
USTC	99.2%	99.7%	99.4%	100%
CTU	98.2%	98.2%	96.3%	98.5%

the most distinguishing pattern and inside, the word "HTTP/" contributes the most for the classification decision. We then identified which traffic feature values are receiving higher attention weights when classified with HAN, and found that more than several site- or communication environment-dependent features, which can not be used as an application traffic signature, had often been considered as discriminative information in making classification decisions; e.g., TCP Window Size, Time-To-Live, Sequence Number, TCP Flags, TCP Options, Header Checksum, Fragment Identification, etc. The other group of features with higher attention weights include port number, packet size info., protocol, number of packets, and payload data, all of which are well-known key features for Internet traffic classification, particularly when used with traditional machine learning algorithms [2] like C4.5.

Table 4: Performance comparisons of DL and ML

Traces	SAE	CNN	LSTM	HAN	C4.5	k-NN (k=3)
WIDE	96.9%	98.7%	99.5%	99.7%	99.7%	87.8%
USTC	53.7%	55.6%	10.1%	89.7%	89.7%	79.8%
CTU	72%	86.9%	80.5%	82.8%	95.9%	87.2%

Next, we re-evaluate the classification performance of the four DL algorithms using the latter group of well-known key features only, excluding payload data, in order to (i) measure how influential those site- or environment-dependent (thus irrelevant for traffic classification) features were in achieving the very high accuracy (>=98.5%) and (ii) compare their performance with those of traditional machine learning algorithms like C4.5 and k-Nearest Neighbors, under the same set of features. As summarized in Table 4, classification accuracy of the DL algorithms significantly drops, particularly in USTC and CTU traces, when the site- or environment-dependent features are excluded. Across all the traces we tested, C4.5 consistently performed the best, achieving the same accuracy with HAN on WIDE and USTC traces. k-NN was the second or third best-performing classifier on CTU and USTC traces, respectively.

Our preliminary results, though still at an early phase, indicate that the high accuracy of all the tested DL algorithms in classifying Internet traffic seems to be an outcome of overfitting to a given trace data, particularly in the process of automatic feature extraction which may often leads to the highest possible accuracy "within' the dataset. As a result, the algorithms choose to automatically extract and use attributes that may look relevant and well-working within a given set of trace, but may not be that useful beyond that, such as TCP Window Size, Time-To-Live, or even IP or MAC addresses when they are not properly masked out. Moreover, when used with the same key well-known traffic features like port number, packet size, protocol, and number of packets, DL-based classifiers show accuracy comparable to or lower than those of traditional machine learning algorithms like C4.5.

Our results make the real benefit and contribution of deep learning over traditional machine learning algorithms to the field of the Internet traffic classification, particularly in building a portable, robust model, still rather questionable. Ongoing work includes further in-depth investigation on this problem, (i) with more recent and diverse DL algorithms like BERT or XLNet, (ii) over more traces with the diverse geographic locations, link characteristics and application traffic mix, and (iii) with more detailed drill-down data analysis.

4 ACKNOWLEDGMENT

This work was supported in part by the National Security Research Institute (NSRI) grant funded by the Korean government (NSRI Institutional Program 2019-014) and in part by the National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIT) (NRF-2019R1A2C1088921). Our special thanks to Chanwoo Bae for motivating us to pursue this study, useful discussion and excellent feedback.

REFERENCES

[1] W Wang et al. 2017. Malware traffic classification using convolutional neural network for representation learning. In *ICOIN*.
[2] Y Lim et al. 2010. Internet traffic classification demystified: on the sources of the discriminative power. In *CoNEXT*.
[3] Z Yang et al. 2016. Hierarchical attention networks for document classification. In *NAACL*.
[4] CTU University. 2016. The Stratosphere IPS Project Dataset. *https://stratosphereips.org/category/dataset.html* (2016).