COMPUTATIONAL INTELLIGENCE AND MACHINE LEARNING

Computational Intelligence and Machine Learning
Vol-3 Issue-2, October 2022, PP.15-21
ISSN : 2582-7464 (Online)

A Review of Deep Learning Techniques for Encrypted Traffic Classification

Auwal Sani Iliyasu [1], Ibrahim Abba [2], Badariyya Sani Iliyasu [3], Abubakar Sadiq Muhammad [4]

[1] Department of Computer Engineering Technology, Kano State Polytechnic, Kano, Nigeria.
[2] Department of Electrical Engineering Technology, Kano State Polytechnic, Kano, Nigeria.
[3] Department of Computer Science, Federal College of Education (Technical) Bichi, Kano, Nigeria.
[4] Department of Computer Engineering Technology, Kano State Polytechnic, Kano, Nigeria.

*Corresponding Author: [1] engrausan@gmail.com, [2] ibrahimaba12950@yahoo.com, [3] ausan84@yahoo.co.uk, [4] alsiddiq4uhd@gmail.com,

Abstract **- Network traffic classification is significant for task such as Quality of Services (QoS) provisioning, resource usage planning, pricing as well as in the context of security such as in Intrusion detection systems. The field has received considerable attention in the industry as well as research communities where approaches such as Port based, Deep packet Inspection (DPI), and Classical machine learning techniques were thoroughly studied. However, the emergence of new applications and encryption protocols as a result of continuous transformation of Internet has led to the rise of new challenges. Recently, researchers have employed deep learning techniques in the domain of network traffic classification in order to leverage the inherent advantages offered by deep learning models such as the ability to capture complex pattern as well as automatic feature learning. This paper reviews deep learning based encrypted traffic classification techniques, as well as highlights the current research gap in the literature.**

Index Terms—Traffic classification, Encrypted traffic, Deep learning, Machine learning

1. Introduction

A factor paramount to the management of a network and its security is Network traffic classification. It makes possible for network operators to apply quality of service (QOS), resource usage planning, pricing and security policies (malware detection, and intrusion detection) in accordance with the need of an application. Recently, the field is attracting more researchers due to rapid changes in technologies associated with the internet and mobile communication. One of the transformations is the use of encryption and obfuscation techniques, which are now prevalent in network applications. Encrypted traffic is known to constitute more than 50% of Internet traffic as a result of increase in demand of privacy from users in order to bypass censorship and enable access to services prohibited geographically [1]. This makes the field of traffic classification to remain active as more challenges arise from use of encrypted network applications.

Traditional approaches to traffic classification are:1) *the port-based approach, 2) payload-based approach and 3) statistical/machine learning approaches*

Port-based Approach: This classification method uses the official Internet Assigned Numbers Authority (IANA) list for classifying applications thereby making it easier for implementation in real time. However, the use of dynamically assigned ports numbers and use of other known port numbers by application to disguise their traffic render this approach ineffective [2].

Payload based approach: This classification method relies on the patterns or keywords in data packet a technique popularly referred to as deep packet inspection (DPI) methods in some literatures. The approach provides very accurate results in classifying applications [3] . Hence, it is mostly employed in commercial tools, however[4], DPI performs poorly against encrypted traffics, and incurs high computational overhead [5].

Statistical/Machine learning-based approaches: To address the limitation of the aforementioned approaches, methods base on flow statistics were introduced. The main intuition behind such methods is that the statistical attributes of network traffic are unique for different applications and can therefore be used to differentiate applications from each other [6]. Machine learning (ML) methods on the other hand are highly efficient in dealing with statistical data[4]. Therefore, several machine learning algorithms such as K-Nearest Neighbor[7], Random Forest [3] , Support Vector[8] Machines are employed.

Computational Intelligence and Machine Learning
Vol-3 Issue-2, October 2022, PP.15-21
ISSN : 2582-7464 (Online)

These methods can effectively classify encrypted traffic, since, they rely on statistical attribute of the network traffic data.

Recently, Deep learning (DL) which is a subset of ML have been widely applied for encrypted traffic classification. These approaches do not require feature engineering, and can automatically learn features from the input data without the need for a domain expert to perform feature selection[9],[10]. Thus, DL techniques perform effectively in large and complex data which characterizes the modern-day network traffic [11]

This work reviews research works that apply Deep learning models to classify encrypted traffic. The paper is organized as follows: Section 2 describes Deep Learning architectures. Section 3 presents our taxonomy of Deep learning-based techniques for encrypted traffic classification. In Section 4, Open challenges are discussed, and finally, section 5 concludes the paper.

2. DL Architectures

DL composed of multiple layers of artificial neurons capable of learning representation/pattern using multiple levels of abstraction. DL has seen considerable adoption in many fields such as computer vision, Natural Language processing etc. This subsection explains some state-of-the-art Deep Learning architectures commonly employed for encrypted traffic classification.

A. Multi-layer Perceptron (MLP)

The Multi-layer Perceptron MLP, also known as feed-forward networks are neural networks architectures with at least one hidden layer beside the conventional input and output layers. Layers in MLP are made up of nodes referred to as neurons, each neuron is fully connected to all neurons in the previous layer. Neurons present in each given layer functions independently without sharing any connection. These layers are connected to provide only unidirectional flow of information. Hence the name feed-forward networks. The primary task of MLP is to approximate any given function by making a neuron takes a sum of dot product of its weights with its inputs, and then pass it through a non-linear activation function to produces an output. The output serves as input to another neuron in the subsequent layer. The last fully connected layer is referred to as the output layer and represents the classes score in the classification context [12].

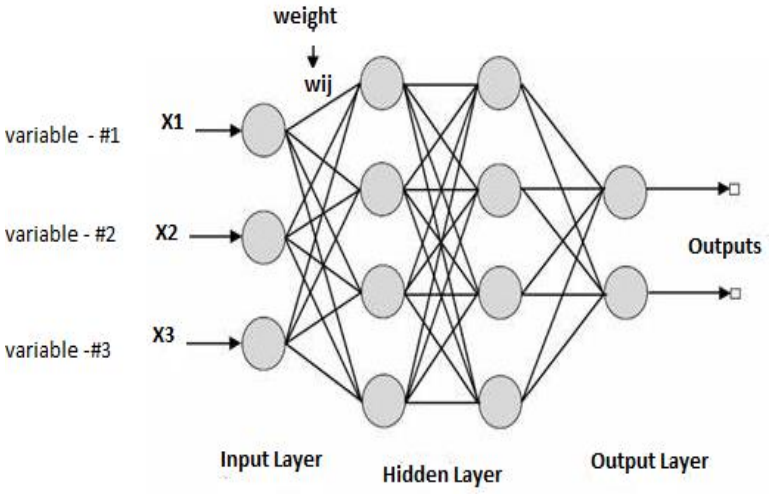

Fig.1. Feed-forward Network Architecture

B. Convolutional Neural Network (CNN)

Convolutional Neural Networks are designed to overcome the drawbacks of overfitting and scaling with high dimensional data associated with regular neural network like MLP, whereby, each neuron in one layer is connected to all other neurons in the next layer. CNN architecture models connectivity pattern of Neurons in mammalian visual cortex, in which individual neurons respond to stimuli only in a limited region of a receptive field. It is made up of a sequence of layers called convolutions. A collection of these fields overlaps to cover the visual area. Each neuron in a convolution layer is connected to a small region of the preceding layer using what is termed as a kernel or filter. This highly reduces the parameter space, and enables it to scales well with data of high dimension. Each layer of a CNN transforms multi-dimensional input volume to another multi-dimensional output volume of neuron activation. However, there exists a layer called pooling which is often sandwiched between one or two convolution layers to enable down sampling of the output. Finally, the last hidden layers of CNN architecture usually employ fully connected layers [13].

Computational Intelligence and Machine Learning
Vol-3 Issue-2, October 2022, PP.15-21
ISSN : 2582-7464 (Online)

C. Recurrent Neural Network (RNN)

A neural network which has a self-recurrent connection in addition to forward flow of information is referred to as Recurrent neural network. It is a form of artificial neural networks in which the self-recurrent connection acts as some kind of memory that allows it to store temporal information. In this architecture, the output of a recurrent neuron at time step t is a function of all the inputs from previous time steps. This feature of the RNN makes it more suited for sequential data such as time series prediction and speech recognition in which good performance has been recorded in a number of literatures. The long-short term memory (LSTM) was introduced to tackle the gradient problem associated with Conventional RNN when training long sequences. LSTM are capable of detecting long-term dependencies in a data and also converge faster. Thus, making them more preferable than the traditional RNN [14].

D. Autoencoder

Another form of artificial neural network is the Autoencoder. This ANN learns to reproduces a given input as its output. The network is composed of Encoder function $h = f(x)$, a feature extraction function which is a hidden layer and a decoder function $r = g(h)$. The internal representation of the input data is learnt by the Encoder function while the decoder function reconstructs the input from the output of the encoder function.

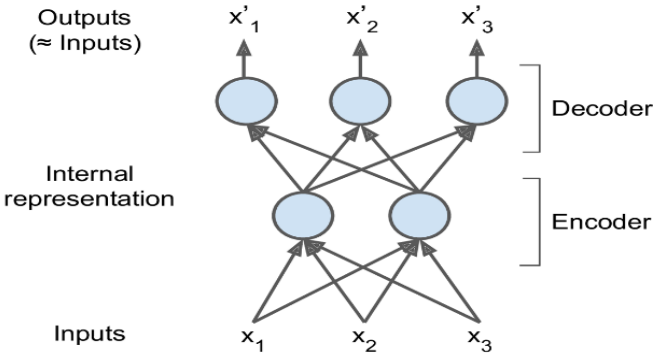

Fig. 3 Autoencoder

It is worth noting that the output is not an exact replica of the input but an approximate value even though the Autoencoder constructs a copy its input as its output. The model is constrained such that it prioritizes which aspect of the input data to learn. As an analogy, noise could be added to the input the network will be trained to recover the original input. The presence of the constraints forces the Autoencoder to learn efficient representation of the input data instead of copying the input directly to the output. This feature makes the Autoencoder suitable for dimensionality reduction as the learned representation which are referred to as codings have much lower dimensionality than the original input data. Autoencoders find suitable application in model where new data that resembles the training data is randomly generated [15].

E. Generative Adversarial Network (GAN)

GAN is a recently developed by Goodfellow et al [citation]. The model comprises of a combination of two neural networks which are trained in adversarial setting. The first network composes of A generator which takes in a random noise and generates new data instances, while the second neural network, receives input from both the generator and the original training data and is termed as discriminator. Each data instances are then reviewed by the discriminator and a decision is made on whether the data is from actual training dataset (real) or from the generator. Theoretically, there exist a point where the generator captures the whole training data distribution and which the discriminator becomes unable to ascertain whether the inputs are from the generator or not. Hence, the GAN is said to be fully trained at this point.

3. Taxonomy of Deep Learning for Encrypted Traffic Classification

This work will employ five criteria to categorize research works in literature that apply deep learning techniques for encrypted traffic classification problem

- Approach
- Features
- Model type
- Classification objective
- Learning mode

Computational Intelligence and Machine Learning
Vol-3 Issue-2, October 2022, PP.15-21
ISSN : 2582-7464 (Online)

A. Approach

This refers to manner the deep learning algorithm is applied to encrypted traffic classification problem. In this category, we have two popular approaches. These are:

- End-to-end approach
- Divide-and Conquer approach

a) End-to-end deep learning approach;

The End-to-end approach leverages the automatic feature learning capability of deep learning models to perform traffic classification. Raw network packet is passed directly to the deep learning model after preprocessing. Thus, the approach eliminates the need of handcrafted features.

b) Divide-and Conquer approach

The divide-and-conquer approach is similar to the way classical machine learning methods are used to handle traffic classification. Therefore, network traffic features are carefully selected, and then feed to the deep learning model. The approach takes advantages of high feature-discriminative ability of deep learning models to offer improve classification accuracy over classical machine learning model.

B. Features

The feature refers to the traffic attribute, which is used as the input to the deep learning model. It mainly comprises the following:

- Packet-based features: this mainly consist of layer 3 and layer 4 header fields such as port numbers, protocols, flags etc. Since, there are several combinations of these fields, the useful ones are carefully selected by domain expert to serve as features to deep learning techniques. However, in a situation where the approach is an and end-to-end, the whole packets can be used as input to the deep learning model
- Flow based features: network flow is described as comprising packets sharing the following five tuples: source IP address, destination IP address, port numbers and protocols. Flow based features comprises mainly of Flow statistics such as minimum packet length, average packet length, volume of packet exchange in forward directions etc. These attributes are obtained after completion or termination of a flow. There exist many combinations of these attributes to be used as features.
- Time-series properties: these are similar to flow-based features; however, time-series features are derived when an arbitrary number of consecutive packets in a given flow are observed instead of an entire flow. The packets can be sampled in any part of a flow not necessarily at the beginning. The features derived may comprises properties such as inter-arrival time between consecutive packets, direction of consecutive packets, packets length etc. One advantage of time-series features over flow-based features is that, they could be used for real-time classification, since features can be generated before completion or termination of a flow. In recent studies [17] where time-series features were employed, as few as 20 packets in a flow were used to achieve a reasonable accuracy.

C. Model type

This refers to the actual deep learning algorithm used in the traffic classification task. Several models and architectures such as MLP, CNN, RNN, LSTM, AE and GAN have been employed. One can refer to Section 2 for a detailed explanation about these models.

D. Classification objective

In this category, models are classified based on the granularity level of their classification task. This mainly falls within the following:

- Binary classification: the main objective here is to classify traffic as belonging to two difference classes. For instance, classifying traffic as either normal or malicious.
- Protocol identification: here traffics are categorized as belonging to a specific protocol. e.g. classifying traffic as HTTP, TCP, TLS, etc.
- Service identification: the task here is to identify a broad category of services to which network traffic belong to. For example, identifying traffic as belonging to streaming services, chat services, etc.
- Application identification: the traffics are tagged as belonging to a specific application. For example Google search, Facebook, YouTube, etc.

Computational Intelligence and Machine Learning
Vol-3 Issue-2, October 2022, PP.15-21
ISSN : 2582-7464 (Online)

E. Learning Mode

The learning mode refers to the way in which the deep learning algorithm is trained. The most popular deep learning mode used in network traffic classification is the supervised learning. However, unsupervised learning and semi-supervised learning methods are also significantly employed.

Table Error! No text of specified style in document.**.1** Taxonomy of recent representative studies of DL techniques in encrypted traffic classification

Ref.	Features		Learning method	Classification goal	Task category
Qing et al. [19]	Flow-based	MLP	Supervised	Coarse-grained	Classifying several protocols
W.Wang et al. [20]	Packet-based	CNN	Supervised	Coarse- grained	classifying vpn Apps.
Lopez et al. [22]	Time-series	LSTM + CNN	Supervised	Coarse-grained	Classifying several protocols
Jonas et al. [24]	Flow-based	AE	Unsupervised	Coarse-grained	Classifying several apps.
Hwang et al. [25]	Packet-based	LSTM+CNN	Supervised	Fine-grained	Classifying several apps.
Shane et al. [26]	Flow-based	MLP	Supervised	Coarse-grained	Protocol identification
Shabaz et al. [27]	Time-series	CNN	Semi-supervised	Coarse-grained	Protocol identification
Iliyasu A.S et al. [17]	Time-series	GAN	Semi-supervised	Coarse-grained	Classifying several apps.

4. Open Challenges

This section highlights some of the open challenges associated with encrypted traffic classification.

A. Large dataset collection and labeling of encrypted traffic

The issue of large data collection and labeling, to apply deep learning models, large and representative dataset is required. The dataset should also be diverse enough to avoid severe overfitting. It is well known that; large dataset collection and labeling is a challenging and non-trivial task. For example, researchers often use DPI (deep packet inspection) tools to label a network flow; however, the proliferation of encryption in today's Internet traffic makes such approach unfeasible. Therefore, labeling is mostly conducted in an isolated environment such as or network edge. One drawback of such approach is model may severely overfit to features peculiar to the user rather than traffic- specific features as the dataset often contains interactions of only one or few users.

B. Multiplex stream

Another issue is that of multiplexed streams where a single flow consists of several traffic classes. This can be pictured in situation where tunneling is in place. The traffic that passes through the tunnel may contain many applications with same source IP address, destination IP address, and protocol and port numbers. One of the challenges is capturing and labeling such traffic. To the best of our knowledge there is no method in the literature that addresses the issue.

5. Conclusion

Network traffic classification serves as the basis for task such network management and security. Several methods have been employed to perform traffic classification. However, emergence of complex challenges such as encryption has paved the way for deep learning models which are better equipped to capture complex patterns than classical machine learning. In this paper, we have reviewed commonly used deep learning models in the domain of network traffic classification, and also

highlighted the current research gap in the literature.

REFERENCES

[1] J. Abbate, "the internet: global evolution and challenges," *THE INTERNET*, p. 9.

[2] N.-F. Huang, G.-Y. Jai, H.-C. Chao, Y.-J. Tzang, and H.-Y. Chang, "Application traffic classification at the early stage by characterizing application rounds," *Information Sciences*, vol. 232, pp. 130–142, May 2013, doi: 10.1016/j.ins.2012.12.039.

[3] Y. Luo, K. Xiang, and S. Li, "Acceleration of decision tree searching for IP traffic classification," in *Proceedings of the 4th ACM/IEEE Symposium on Architectures for Networking and Communications Systems - ANCS '08*, San Jose, California, 2008, p. 40. doi: 10.1145/1477942.1477949.

[4] T. Seyed Tabatabaei, M. Adel, F. Karray, and M. Kamel, "Machine Learning-Based Classification of Encrypted Internet Traffic," in *Machine Learning and Data Mining in Pattern Recognition*, vol. 7376, P. Perner, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 578–592. doi: 10.1007/978-3-642-31537-4_45.

[5] P. Velan, M. Čermák, P. Čeleda, and M. Drašar, "A survey of methods for encrypted traffic classification and analysis: a survey of methods for encrypted traffic classification and analysis," *Int. J. Network Mgmt*, vol. 25, no. 5, pp. 355–374, Sep. 2015, doi: 10.1002/nem.1901.

[6] T. Bujlow, "Classification and Analysis of Computer Network Traffic," p. 288.

[7] D. J. Arndt and A. N. Zincir-Heywood, "A Comparison of three machine learning techniques for encrypted network traffic analysis," in *2011 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA)*, Paris, France, Apr. 2011, pp. 107–114. doi: 10.1109/CISDA.2011.5945941.

[8] A. Este, F. Gringoli, and L. Salgarelli, "Support Vector Machines for TCP traffic classification," *Computer Networks*, vol. 53, no. 14, pp. 2476–2490, Sep. 2009, doi: 10.1016/j.comnet.2009.05.003.

[9] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. Cambridge, Massachusetts: The MIT Press, 2016.

[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[11] M. Abbasi, A. Shahraki, and A. Taherkordi, "Deep Learning for Network Traffic Monitoring and Analysis (NTMA): A Survey," *Computer Communications*, vol. 170, pp. 19–41, Mar. 2021, doi: 10.1016/j.comcom.2021.01.021.

[12] Xue-Wen Chen and Xiaotong Lin, "Big Data Deep Learning: Challenges and Perspectives," *IEEE Access*, vol. 2, pp. 514–525, 2014, doi: 10.1109/ACCESS.2014.2325029.

[13] K. O'Shea and R. Nash, "An Introduction to Convolutional Neural Networks," *arXiv:1511.08458 [cs]*, Dec. 2015, Accessed: Apr. 09, 2022. [Online]. Available: http://arxiv.org/abs/1511.08458

[14] A. Graves, "Generating Sequences With Recurrent Neural Networks," *arXiv:1308.0850 [cs]*, Jun. 2014, Accessed: Apr. 09, 2022. [Online]. Available: http://arxiv.org/abs/1308.0850

[15] S. R. Jurie Fr´ed´eric, "Discriminative Autoencoders for Small Targets Detection.," *Razakarivony, Sebastien, and Frédéric Jurie. "Discriminative autoencoders for small targets detection." 2014 22nd International conference on pattern recognition. IEEE, 2014.*.

[16] I. Goodfellow *et al.*, "Generative Adversarial Nets," p. 9.

[17] A. S. Iliyasu and H. Deng, "Semi-Supervised Encrypted Traffic Classification With Deep Convolutional Generative Adversarial Networks," *IEEE Access*, vol. 8, pp. 118–126, 2020, doi: 10.1109/ACCESS.2019.2962106.

[18] Z. Wang, "The Applications of Deep Learning on Traffic Identification," p. 10.

[19] Q. Lyu and X. Lu, "Effective Media Traffic Classification Using Deep Learning," in *Proceedings of the 2019 3rd International Conference on Compute and Data Analysis*, Kahului HI USA, Mar. 2019, pp. 139–146. doi: 10.1145/3314545.3316278.

[20] W. Wang, M. Zhu, J. Wang, X. Zeng, and Z. Yang, "End-to-end encrypted traffic classification with one-dimensional convolution neural networks," in *2017 IEEE International Conference on Intelligence and Security Informatics (ISI)*, Beijing, China, Jul. 2017, pp. 43–48. doi: 10.1109/ISI.2017.8004872.

[21] Z. Chen, K. He, J. Li, and Y. Geng, "Seq2Img: A sequence-to-image based approach towards IP traffic classification using convolutional neural networks," in *2017 IEEE International Conference on Big Data (Big Data)*, Boston, MA, Dec. 2017, pp. 1271–1276. doi: 10.1109/BigData.2017.8258054.

[22] M. Lopez-Martin, B. Carro, A. Sanchez-Esguevillas, and J. Lloret, "Network Traffic Classifier With Convolutional and Recurrent Neural Networks for Internet of Things," *IEEE Access*, vol. 5, pp. 18042–18050, 2017, doi: 10.1109/ACCESS.2017.2747560.

[23] O. Aouedi, K. Piamrat, and D. Bagadthey, "A Semi-supervised Stacked Autoencoder Approach for Network Traffic Classification," in *2020 IEEE 28th International Conference on Network Protocols (ICNP)*, Madrid, Spain, Oct. 2020, pp. 1–6. doi: 10.1109/ICNP49622.2020.9259390.

[24] J. Hochst, L. Baumgartner, M. Hollick, and B. Freisleben, "Unsupervised Traffic Flow Classification Using a Neural Autoencoder," in *2017 IEEE 42nd Conference on Local Computer Networks (LCN)*, Singapore, Oct. 2017, pp. 523–526. doi: 10.1109/LCN.2017.57.

[25] R.-H. Hwang, M.-C. Peng, V.-L. Nguyen, and Y.-L. Chang, "An LSTM-Based Deep Learning Approach for Classifying Malicious Traffic at the Packet Level," *Applied Sciences*, vol. 9, no. 16, p. 3414, Aug. 2019, doi: 10.3390/app9163414.

[26] P. Wang, F. Ye, X. Chen, and Y. Qian, "Datanet: Deep Learning Based Encrypted Network Traffic Classification in SDN Home Gateway," *IEEE Access*, vol. 6, pp. 55380–55391, 2018, doi: 10.1109/ACCESS.2018.2872430.

[27] S. Rezaei and X. Liu, "How to Achieve High Classification Accuracy with Just a Few Labels: A Semi-supervised Approach Using Sampled Packets," *arXiv:1812.09761 [cs]*, Dec. 2018, Accessed: Aug. 17, 2019. [Online]. Available: http://arxiv.org/abs/1812.09761

[28] M. Yeo *et al.*, "Flow-based malware detection using convolutional neural network," in *2018 International Conference on Information Networking (ICOIN)*, Chiang Mai, Jan. 2018, pp. 910–913. doi: 10.1109/ICOIN.2018.8343255.

[29] L. Vu, C. T. Bui, and Q. U. Nguyen, "A Deep Learning Based Method for Handling Imbalanced Problem in Network Traffic Classification," in *Proceedings of the Eighth International Symposium on Information and Communication Technology*, Nha Trang City Viet Nam, Dec. 2017, pp. 333–339. doi: 10.1145/3155133.3155175.