Deep Learning for Network Traffic Classification: Methods, Datasets, and Future Directions

Priya Gupta

Campbell University

Priyapp@gmail.com

Abstract:

With the rapid advancement of network technology and the growing reliance on internet services across various sectors, effective network traffic management has become imperative. Network traffic classification plays a crucial role in this management by categorizing and predicting network traffic, thereby aiding in quality control, custom pricing, resource allocation, and testing. Traditional methods such as port-based and signature-based classification have become less effective due to evolving application techniques. Deep learning, a subset of machine learning, offers a robust solution by training classification models without explicit programming and handling complex patterns efficiently.This paper delves into network traffic identification using deep learning techniques, providing an overview of existing methods and analyzing both machine learning and deep learning approaches. A comprehensive examination of relevant literature and datasets is conducted, highlighting key representative datasets used in deep learning-based traffic identification. The study categorizes and evaluates deep learning methods, including multi-layer perceptron, convolutional neural network (CNN), and recurrent neural network (RNN), discussing their respective advantages and limitations. Special attention is given to the shortcomings of CNNs in traffic identification, proposing the use of attention mechanism-based CNNs for improved context recognition. Finally, the paper outlines potential future applications and advancements in deep learning methods for network traffic identification.

Keywords:

Network traffic; Traffic identification; deep learning; trends.

1. Introduction

With the continuous progress of network technology and the update and iteration of various network service products, the Internet has promoted the development of global communication. In recent years, all walks of life, including business, education and individuals, have shown unprecedented dependence on the Internet. Because of this dependence, coupled with the exponential growth of network traffic generated, it is crucial to effectively manage the basic network, and the social demand for network management technology is also very urgent[1].

Network traffic classification is an important part of network management, mainly in the recognition through the type and amount of network traffic, it is the function of network traffic can be divided into appropriate categories, and allow the forecast of traffic on the network, it is very important for many applications, such as service quality control, custom price, resource allocation management, part of the test, and so on. Due to its importance, researchers have developed many different methods over the years to adapt to the diverse and changing needs of application scenarios. The development of Internet benefits has improved users' sense of use. The network traffic classification method is divided into port-based method and signature- based method. Port-based classification operates by classifying network traffic according to the port number contained in the header section of each datagram sent over the network. This simplified approach has become ineffective in recent years because it simplifies ways that applications can bypass by changing port Numbers. The signature-based approach involves matching a sequence of bytes or attributes with a known signature, but is limited to identifying the signature of a known application protocol.

Deep learning is a subfield of machine learning that enables training of classification models without explicit programming. Deep learning models usually contain a large number of processing units that can be used to classify data. Deep learning eliminates the need for domain expert selection by training the automatic selection function, which makes deep learning an ideal method for traffic identification. Another important feature of deep learning is that it is more capable of learning than traditional machine learning methods, so it can learn highly complex patterns. It is not necessary to decompose the problem into a smaller problem of feature selection and classification [3].

2. Research Status at Home and Abroad

Many existing network traffic identification methods, especially machine learning and deep learning methods, do not need to rely on port number and packet load information, and are not affected by the encryption load. Combined with machine learning and deep learning technology, the statistical law of flow can be analyzed more effectively to classify traffic. Among them, the statistical features of flow mainly include the size and number of packets, etc. Moore et al. [4] proposed 248 features of flow, so most researchers chose to take a certain subset of features on this basis for their own relevant research work. In 2003, researchers [5] proved for the first time the correspondence between packet length and network protocol, which can be used for network protocol identification. In 2005, Zuev et al. [6] adopted the Naive Bayes method in the machine learning algorithm to extract certain network traffic characteristics for training, and finally achieved the recognition effect. However, the classification accuracy of this method was only about 60%. Then, Huang [7] KNN algorithm of machine learning is used to classify the flow of network application research, finally studies the experimental results show that the classification accuracy of the method about 90%, but this method has certain defects, is there is a packet arrival, algorithm for training on all flow calculation, a new the repeated calculation formula makes the classification performance of the algorithm is low. In 2009, Xu Peng et al. [8] used The C4.5 decision tree in the algorithm to classify network applications, and the experiment showed that the classification accuracy of this method could reach 94%. However, the classification method of C4.5 decision tree requires more flow characteristics and data grouping, and the computational complexity of the algorithm is also high. In 2014, Patel et al.

[9] adopted an iterative SVM algorithm for P2P application, and the classification accuracy rate was only about 70%. In 2015, Hong et al. [10] used SVM algorithm to classify applications with an accuracy rate of only about 80%. This shows the currently existing web application classification method usually has certain limitation, the network traffic classification method based on machine learning, although has a good classification effect, but its actual performance is not so stable recognition accuracy, and has high computing complexity, will be the management of network traffic and network application category of uncontrollable factors.

3. Identify Data Sets based on Deep Learning Network Traffic

Many studies adopt manual collection or private data sets of companies, which affects the credibility of conclusions to a certain extent. After the research and development in recent years, there have been a lot of public data sets, this paper will select a strong representative of the data set to introduce.

3.1. ISCXVPN2016 Data Set

The ISCXVPN2016 dataset divides network traffic into traffic passing through different types of ports and traffic passing through different types of applications.

Among them, the types of port traffic through different types mainly include regular session traffic and VPN session traffic, which can be used to effectively identify VPN traffic and non- VPN traffic. The traffic through different types of applications is mainly divided into browsing traffic, email traffic, chat traffic, media traffic, file transfer traffic, VoIP traffic and TraP2P.

3.2. Moore Data Set

Moore open DATA set was created by Professor AndrewW. Moore Laboratory, Cambridge University. According to different applications, the DATA set divides network traffic into 12 categories, including WWW, MAIL, FTP-Conrol, FTP-PASV, ATTACK, P2P, DATABASE, FTP-Data,MULTIMEDIA, SERVICES, INTERACTWE and GAMES.

3.3. USTC-TFC2016 Data Set

USTC - TFC2016 datasets published by university of science and technology of China in 2016, the data set is mainly made up of two parts: one is from CTU (the Czech technical university) selection of 10 kinds of malicious traffic network data set, they are made by researchers at the university of CTU in 2011 to 2015 collected from the real world, and for some size larger files for the interception, merged for small size file processing. The other is 10 kinds of normal traffic newly collected by IXIABPS, a professional network traffic simulation device developed by IXIA.

4. Traffic Identification Algorithm based on Deep Learning Network

After a lot of experiments in the development of deep learning, the existing network traffic recognition algorithms based on deep learning are mainly divided into the recognition algorithm based on multi-layer perceptron, the recognition algorithm based on convolutional neural network and the recognition algorithm based on recursive neural network. This chapter will give a brief introduction to these three methods.

4.1. Multilayer Perceptron

Multi-layer perceptron (MLP) is the first neural network architecture consisting of input layer, output layer and several hidden layers of neurons. Each layer has several neurons, tightly connected to adjacent layers. Neurons obtain the weighted sum of their inputs and generate outputs through nonlinear activation functions. Theoretically, a dense and deep enough MLP can estimate any arbitrary function. However, because the model needs to learn a large number of parameters, it is often very complex, inefficient, and difficult to train for arbitrarily complex problems. Although the use of only deep MLPS has been rejected, several layers of fully connected neurons can be used as a small part of other models.

The BP algorithm of MLP is based on the classical derivative rule of chain. Firstly, look at forward conduction. For the input layer, there are I units; for the input sample (x,z), the input of the hidden layer is:

$$a_h = \sum_{i=1}^{I} w_{ih} x_i$$

$$b_h = f(a_h)$$

After computing the conduction from the input layer to the first hidden layer, the remaining hidden layer is calculated in a similar way. H_l is used to represent the number of units in the L layer:

$$a_h = \sum_{h'=1}^{h_{l-1}} w_{h'h} b_{h'}$$

$$b_h = f(a_h)$$

For network traffic classification, pure multilayer perceptron (MLP) is rarely used due to its complexity and low accuracy. G. Aceto[12] used three mobile data sets with different number of tags to compare the deep learning method with the random forest (RF) algorithm to show the performance gap. The performance of deep learning method is higher than RF. However, because RF, MLP, and other deep learning methods use different input functions, the experimental setup is not entirely fair and equal. Therefore, the results should not be considered a comprehensive comparison of ML methods.

4.2. Convolutional Neural Network

The high-dimensional input does not work well, resulting in a large number of learnable parameters in the hidden layer, which the CNN architecture solves by using the convolutional layer. Using these kernels over the entire input also helps the model to capture displacement invariants more easily. Pooled layers are also used for secondary sampling after one or more convolutional layers, and fully connected layers are usually used for the last hidden layer.

In the process of processing, matrix convolution is often used to calculate image features. Matrix convolution is divided into full convolution and effective value convolution. The formula of full convolution is as follows:

$$z(u,v)=\sum_{i=-\infty}^{\infty}\ \sum_{j=-\infty}^{\infty}\ x_{i,j}\cdot k_{u-i,v-j} \tag{1}$$

Assuming that X is an m* M order matrix, K is an N *n order matrix, and Krot is obtained by rotating K 180 degrees, then the effective value convolution formula is:

$$\begin{gathered} z(u,v)=\sum_{i=-\infty}^{\infty}\ \sum_{j=-\infty}^{\infty}\ x_{i+u,j+v}\cdot k_{roti,j}\cdot\chi(i,j) \\ \chi(i,j)=\begin{matrix}1,0\leqslant i,j\leqslant n\\ 0,\text{others}\end{matrix} \end{gathered} \tag{2}$$

W. Wang [13] proposed a convolutional neural network (CNN) model with the simplest method, which basically represented each stream or session with a 1-dimensional vector, and the evaluation results showed a significant improvement over the C4.5 method using time series and statistical functions. Z. Chen, [14] will CNN with two convolution, two pools and three fullyconnection layer is used to deal with and application classification task, and use the reproducing kernel Hilbert space (RKHS) embedded, the early time series data is converted into a 2 d image, the proposed CNN model in terms of agreement and application classification taskis better than the traditional ML method and MLP. The architecture of the convolutional neural network is shown as follows:

4.3. Recursive Neural Network

A recursive neural network (RNN) is a neural network that contains loops to store time information. Recursive neural networks are designed specifically for sequential data, where the output has a final input and a previous input decision. Recursive neural networks have been successfully applied to speech recognition, time budgeting, translation and language modeling. Gradient disappearance and explosions, which make it difficult to learn long-term dependencies (for example, dependencies between inputs), are common obstacles in traditional RNN. Long-term short-term memory (LSTM) is introduced to alleviate these problems by adding a set of gates that control when information is stored or deleted. The schematic diagram of recurrent neural network is as follows:

As an important part of the recursive neural network, it can be represented by an internal hidden state H, which is updated at each time step. The function of this updating mode is as follows:

$$\left\{\begin{matrix} h_t=f(W_{xh}x_t+W_{hh}h_{t-1}+b_1) \\ y_t=g(W_{hy}h_t+b_2) \end{matrix}\right. \tag{3}$$

Where x_t and W_xh are the input value and weight matrix respectively, f and g are activation functions, and b_1 and b_2 represent the deviation between the output and the original value. For network classification tasks, according to relevant studies, the hybrid model is superior to the pure LSTM or CNN model. In order to capture the spatial and temporal characteristics of the flow, W. Wang[15] and M. Lopez-Martin[16] used BOTH CNN and RNN for different application classifications. In addition to slight differences, the contents of the first 6 to 30 packets were fed into the CNN model, and then input into the RNN or LSTM model. The fixed input mechanism has high accuracy under different neural network architectures and data sets. Although the LSTM has been successful with sequential data, they are not suitable for complex tasks that require explicit and external storage. New architectures, such as memory networks and neural Turing machines (NTM), have recently been introduced to embed explicit memory in the architecture, known as memory enhanced neural networks (MANN). MANN has been successfully applied to language modeling, question answering and single learning. MANN's performance on network classification task has not been studied.

5. Traffic Identification Experiment based on Convolutional Neural Network

5.1. Experimental Environment and Data

The experimental data set was formed in a network consisting of 8 computers, including 3 Win7 hosts, 3 Ubuntu system hosts, 1 Windows XP host, and 1 Ubuntu server. The data tries its best to imitate the network environment closest to daily life, and also includes common application layer communication protocols, such as HTTP and DNS. The capture of the data set is generated by the VBS toolkit, along with the process name, quintuple, and other related information.

In the whole data set, there are a total of 737,681 data streams, 96.36% of which are labeled as application layer protocols, 95.13% are labeled as application layer protocols, among which two-thirds of each application is used as training data and the rest as test data. The evaluation criterion of traffic identification method is the rate of network traffic detection.

5.2. Experimental Process

The convolutional neural network model mentioned above trains and classifies data sets at the application category level. Relu function is used for activation function, flexible maximum transfer function is used for classifier. The whole connectivity network is the output layer and the previous layer, and stochastic gradient descent algorithm is used for optimization of loss function. The structure of the convolutional neural network refers to its application in handwriting recognition. There are two convolutional layers and one sampling layer. Finally, the flexible maximum transfer function classifier is connected in a fully connected way.

6. Introduction of Future Research Methods and Application Fields

According to the current development trend of deep learning, there will be more methods applied in network traffic identification in the future. This chapter will expand the network traffic identification methods and applications based on deep learning.

6.1. Multi-label Classification

A single stream may contain more than one class tag, called a multiplexed stream. For example, traffic passing through the tunnel may contain multiple applications sharing the same quad, the QUIC protocol may also contain several types of traffic, and there is no way in the traffic classification or related literature to handle these situations. The most difficult challenge is how to properly collect and tag such traffic.

6.2. Mid-stream Identification

About 90% of current researchers' traffic is transient, and in some applications, such as the ASTAF project, they may want to focus on the long process. However, if the classification method relies on the first few packets, the ISP should store the first few packets of all traffic, which is a huge burden. On the other hand, if the classification method is suitable for packets in the middle of the stream, the ISP can wait and detect the elephant stream and then classify the elephant stream by capturing some packets from the middle of the stream. This will greatly reduce memory and computation overhead. Some studies have shown greater accuracy when it comes to the first few packets, but no comprehensive study has been done to use a set of packets at any point in the stream. Some studies divide the whole traffic into several bursts, and then classify each burst to detect different user operations [14]. This means that the onset of a burst should also be detected, and the capture process must begin at that particular point. In addition, it is not clear whether this approach applies to other classification problems other than user operations. In [15], the author USES a fixed number of sampled packets from different parts of the stream for classification. They are moderately accurate when sampled from anywhere in the flow, but the high precision from the intermediate flow is still an unresolved issue.

6.3. Zero Time Difference Applications

A zero time difference application is a new transportation category whose sample does not exist in the training set. Studies have shown that in some cases zero-time applications can account for 60 percent of network traffic and 30 percent of bytes. Despite its significance, it is still in its infancy, and only a few recent studies [16] have proposed solutions that usually rely on detecting and then labeling unlabeled clusters. In recent studies on the classification of character images [21], a combination of reinforcement learning and LSTM has been used to perform one of two possible actions: predicting categories or seeking new tags.

6.4. Transfer Learning or Domain Adaptation Learning

Migration learning allows the model trained on the source task to be used for other target tasks. This process is only effective if the model learning function is not specific to the source task. Because the model has been trained to capture useful features, the retraining process for the target task requires significantly less tagging data and training time. In the case of network traffic classification, the model can be pre-trained using publicly available data sets and then adjusted to another traffic classification task with fewer marker samples.

Although it is different from migration learning, similar techniques have been used to address both problems. In the case of network traffic classification, an example is to train the traffic classifier model using the data set captured on the communication client, and then to classify traffic for network cores with different data distributions. As another example, the model can be periodically retrained based on domain adaptive techniques to capture new patterns of classes with changing characteristics, which is not uncommon in today's networks. While they are useful, these policies have not been widely used for network classification tasks.

6.5. Multitask Learning

The method refers to any model in which multiple loss functions are optimized. A typical approach is to share a hidden layer among all tasks, and each task has its own output layer. Studies have shown that it reduces the risk of overfitting and helps models find relevant features more quickly. This helps when input data is generated from similar probability distributions, or when a set of transformations can be used to generate each other. Thus, if they are similar to the target task data set, it is possible to use other available data sets and define a task for each task, which can easily expand the data set and improve generalization. Many variations of multitask learning have been successfully used in natural language processing and computer vision. Experiments have shown that adding some auxiliary tasks can improve generalization and performance even for single-task problems, but no research has been conducted on network traffic classification tasks.

7. Summarize

In this paper, the network traffic identification based on deep learning method is studied, the existing methods and contents of network traffic identification are understood, and the existing machine learning and deep learning methods are roughly analyzed and elaborated. In order to further analyze network identification methods based on deep learning, this paper, through the study of relevant literature, understands the existing traffic identification data sets based on deep learning network, and introduces the representative data sets among them. At the same time, according to the existing categories of deep learning methods, the network traffic identification methods based on deep learning are classified and analyzed respectively from multi-layer perceptron, convolutional neural network and recursive neural network, and the advantages and disadvantages of these five methods are roughly analyzed. Through further research, this paper analyzes the defects of convolutional neural network in network traffic identification, and considers the use of CNN network based on attention mechanism to identify traffic context content. Finally, based on the above research and the development status of deep learning, this paper proposes relevant deep learning methods and application fields that may be used in network recognition in the future.

References

[1] Smit D, Millar K, Page C, et al. Looking deeper: Using deep learning to identify Internet communications traffic[C]. Australasian Conference of Undergraduate Research (ACUR), Adelaide. 2017.

[2] Rezaei S, Liu X. Deep learning for encrypted traffic classification: An overview[J]. IEEE communications magazine, 2019, 57(5): 76-81.

[3] Zuev D, Moore A W. Traffic Classification Using a Statistical Approach[C]. Passive & Active Network Measurement, International Workshop, Pam, Boston, Ma, Usa, March 31-april. Springer-Verlag, 2005.

[4] Huang S, Chen K, Liu C, et al. A statistical-feature-based approach to internet traffic classification using machine learning[C]. 2009 International Conference on Ultra Modern Telecommunications & Workshops. IEEE, 2009: 1-6.

[5] O'Connor P, Neil D, Liu S C, et al. Real-time classification and sensor fusion with a spiking deep belief network[J]. Frontiers in neuroscience, 2013, 7: 178.

[6] Tsang I W, Kwok J T, Cheung P M. Core vector machines: Fast SVM training on very large data sets[J]. Journal of Machine Learning Research, 2005, 6(Apr): 363-392.

[7] G. Aceto, C. Domenico, M. Antonio, and P[J]. Antonio, ”Mobile encrypted traffic classification using deep learning,” In 2018 Network Traffic Measurement and Analysis Conference (TMA), Jun. 2018, pp. 1-8.

[8] W. Wang, M. Zhu, J. Wang, X. Zeng, and Z[J]. Yang, ”End-to-end encrypted traffic classification with one-dimensional convolution neural networks,” In Intelligence and Security Informatics (ISI), IEEE International Conference on. IEEE, Jul. 2017, pp. 43-48.

[9] Z. Chen, K. He, J[J]. Li, and Y. Geng, ”Seq2Img: A sequence-to-image based approach towards IP traffic classification using convolutional neural networks,” Big Data (Big Data), 2017 IEEE International Conference on. IEEE, 2017, pp. 1271-1276.

[10]W. Wang, et al, ”HAST-IDS: learning hierarchical spatial-temporal features using deep neural networks to improve intrusion detection,” IEEE Access, vol[J]. 6, 2018, pp. 1792-1806.

[11]M. Lopez-Martin, B. Carro, A. Sanchez-Esguevillas, and J[J]. Lloret, ”Network traffic classifier with convolutional and recurrent neural networks for Internet of Things,” IEEE Access, vol. 5, 2017, pp.18042-18050.

[12]J. Hochst, L[J]. Baumgartner, M. Hollick, and B. Freisleben, ”Unsupervised Traffic Flow Classification Using a Neural Autoencoder,” 2017 IEEE 42nd Conference on Local Computer Networks (LCN), IEEE, Oct. 2017, pp. 523-526.

[13]V. F[J]. Taylor, R. Spolaor, M. Conti, and I. Martinovic, ”Robust smartphone app identification via encrypted network traffic analysis,” IEEE Transactions on Information Forensics and Security, vol. 13, no. 1, Jan. 2018, pp. 63-78.

[14]S. Rezaei, X[J]. Liu, ”How to Achieve High Classification Accuracy with Just a Few Labels: A Semi-

supervised Approach Using Sampled Packets,” arXiv preprint arXiv:1812.09761 (2018).

[15]J. Zhang, et al[J]. ”Robust network traffic classification,” IEEE/ACM Transactions on Networking (TON), vol. 23, no. 4, 2015, pp. 1257-1270.

[16]M. Woodward, and C[J]. Finn. ”Active one-shot learning.” In NIPS (2016) Deep Reinforcement Learning Workshop, 2018.M. Lotfollahi, R. Shirali, M.J. Siavoshani, and M. Saberian, ”Deep Packet: A Novel Approach For Encrypted Traffic Classification Using Deep Learning,” arXiv preprint arXiv:1709.02656 (2017).