Workgroup: Network Management
Internet-Draft: draft-gww-mten-dfi-00
Published: 18 August 2022
Intended Status: Standards Track
Expires: 19 February 2023

Authors: Y. Lei *Huawei*, J. Wu *Huawei*, X. Sun *Huawei*, L. Zhang *Huawei*, Q. Wu *Huawei*

Encrypted Traffic Classification Through Deep Learning

Abstract

Quickly and accurately classify applications is important for network congestion control and network service assurance. However with the increased usage of data encryption, privacy enhancing technologies, it became difficult to obtain metadata or sample labels for private enterprise applications. This position papers discusses encrypted traffic classification through deep learning technology. A flow-based classification mechanism is proposed, which only relies on the statistical characteristics of packets, such as time series characteristics, 5 tuple information for feature extraction.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

The popularity of networks and the diversification of services make the internet traffic surge, as a result, the network congestion occurs, which would increases forwarding delay, and even loss packets. The best solution to the problem of network congestion is to increase the network bandwidth, but at the cost of operation and maintenance, which is unrealistic. The most cost-effective solution is to apply a "guaranteed" strategy to manage the network traffic on demand, and prioritize service quality for high-value and critical businesses. For example, the intelligent routing capability of SD-WAN can realize that low-priority applications make way for high-priority applications when congestion occurs to ensure the network experience of key applications. And the base technology of SD-WAN intelligent routing capability is Quality of service (QoS) technology, which allows different traffic to compete unequally for limited network resources, so that voice, video and important data applications can be preferentially served in network devices. The mainstream QoS service model - Differentiated Services model is based on packet classification and marking. Therefore, how to quickly and accurately classify applications is the key to solve network congestion and ensure network service quality.

The earliest network traffic classification method is to use the transport layer protocol UDP or TCP port number for classification, this method is easy to implement, and has low time complexity. But with the diversification of applications and protocols, and the development of port hopping and port masquerading technology appears, the accuracy of the traffic classification method based on port identification is decreasing, so the method is no longer reliable. Therefore, the Deep Packet Inspection (DPI) technology [DPI]which based on parsing the packets'payload, has gradually attracted more attention. DPI technology mainly analyzes the payload of the data packet in the network flow. If the payload part can match the currently known application or protocol with some characteristics, then it can be considered with high probability that this network flow is the known application or protocol. However, due to increased usage of data encryption, privacy enhanced technologies, and the difficulty of obtaining sample labels for private enterprise applications, it is unlikely to get this technology widely deployed. Therefore, a flow-based classification algorithm is needed, which only relies on the statistical characteristics of packets, such as time series characteristics, 5 tuple information in the session, there is no need to parse packets, thus avoiding user privacy issues.

2. Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119][RFC8174] when, and only when, they appear in all capitals, as shown here.

3. DFI-based application identification

Deep Flow Inspection(DFI)[ETC] is an application identification technology based on traffic behavior. It implements application classification by mining the data flows'packet statistical information, the timing relationship between packets, and the association relationship between sessions. Since it doesn't need to obtain the packets'payload information, it does't involve security and privacy issues. At the same time, data encryption does not change the packet statistics distribution and packet timing characteristics, so the classification accuracy of the model is not affected by data encryption.

However, the deep learning classification algorithm based on traffic behavior faces the following problems: more classifications leads to lower accuracy, unbalanced number of classification samples, high misidentification rate of unknown applications or new applications.

In order to solve the problem that the recognition accuracy decreases and the model expands sharply due to multiple classifications, based on the principles of QoS technology, we consider classifying actual applications by application category. We classify applications into voice, video, file transfer and other application categories. This reduces the model size and improves the model classification efficiency. In addition, each application category has a corresponding top application. Thus, the problem of imbalance in the number of classification samples can be limited to imbalances between application categories. Aiming at the problem of endless application updates, enumerated unknown classification samples, which leads to the model identification accuracy decreasing and the difficulty of engineering the algorithm, we could solve it by adding "Other" labels to identify applications that are not in known categories as "Other" applications, which also improves the robustness of the model.

4. DFI-based application Identification process

Figure 1 shows the identification process. Different types of applications present different statistical features and sequence features in the flow direction and the flow length. The classification system calculates the distribution of the first N packets per flow, and then divides the samples into known applications and unknown applications through unsupervised models. Applications of known categories continue to use deep neural models to further mine the statistical and time series features of flows, and mine frequent time series relationships between flows to form a time series combination pattern of multiple flows (e.g. game action streams and heartbeat streams, and concurrent download streams). Finally, applications are classified into voice, video, game, file transfer, etc. Based on QoS policies, different QoS service assurance is provided for different valued services, that is, different bandwidth, delay, packet loss, and congestion. In addition, unknown application categories can be cached as required, and then periodically analyzed and labeled to classify them into known application categories. In this way, a new sample is generated to update the old model, so that the updated model can correctly classify the features of the old unknown application categories.

```
+---------------------+
|     /---------\     |               +--------------------------------+
|  ///           \\\  | ++    ++      |                                |
| |  Known Apps     | | ||++  ||      | Unsupervised    CNN Model      |
||   Wechat, Youtube || ||||++||      | Learning            |          |
| |  SAP, SKYPE     | | ++++++++      |     |               |          |
|  \\\           ///  |               |+----V---------++----V---------+|
|     \---------/     |-------------- || Unknown      ||Known         ||
|                     |  N packets    || Application  ||Application   ||
|   +---------------+ |  flow         ||Identification||Identification||
|   |  Unknown Apps | |  direction    |+--+-----------++-----+--------+|
|   |  Bittorent    | |  and Length   +---+-----------------+---------+
|   +---------------+ |                   |                  |
+---------------------+                   |                  |
                                       +--+-----+       +----+---+
                                       |VPN     |       |Audio   |
                                       |TOR     |       |Video   |
                                       |Other   |       |Game    |
                                       +--------+       +--------+
```

In the traffic identification algorithm, the deep learning model can be selected freely, such as CNN and RNN. LeCun [DL] shows that CNN is suitable for data in the following format: data is presented in array format, data has strong local correlation, and features can appear anywhere in the data. Therefore, in this identification framework, considering the characteristics of the stream and the feature format extracted from the stream, we choose the CNN algorithm which is commonly used in the field of Machine Vision. For the training sample imbalance problem described above, we limit the imbalance problem to the imbalance between different application categories through application categories classification. Here, we solve the data imbalance problem between application categories by optimizing the cost function of CNN, for example, increasing the weight of subclasses. In this way, the issue of model bias caused by unbalanced sample size is mitigated greatly or resolved.

5. Conclusion

With the continuous development of encryption protocols, Internet data transmission enters into the era of full encryption, which poses a challenge to network management. In this paper, based on QoS characteristics in the encrypted traffic management, an encryption traffic classification method is proposed. Through accurate and fast classification of encrypted traffic, the specified application categories'QoS (packet loss, delay, jitter) is guaranteed.

6. Informative References

[DL] LeCun, Y., Bengio, Y., Hinton, G., "Deep learning", Nature vol. 521, pp. 436-444, May 2015.

[DPI] Dharmapurikar, S., Krishnamurthy, P., Sproull, T., "Deep packet inspection using parallel bloom filters", IEEE 11th Symposium on High Performance Interconnects, 2003. Proceedings, 2003.

[ETC] Aceto, G., Ciuonzo, D., Montieri, A., "Toward effective mobile encrypted traffic classification through deep learning", Neurocomputing vol. 409, pp. 306 - 315, 2020.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

Authors' Addresses

Yupeng Lei
Huawei
101 Software Avenue, Yuhua District
Nanjing
Jiangsu, 210012
China
Email: leiyupeng1@huawei.com

Jun Wu
Huawei
101 Software Avenue, Yuhua District
Nanjing
Jiangsu, 210012
China
Email: junwu.wu@huawei.com

Xudong Sun
Huawei
101 Software Avenue, Yuhua District
Nanjing
Jiangsu, 210012
China
Email: sunxudong1@huawei.com

Liang Zhang
Huawei
101 Software Avenue, Yuhua District
Nanjing
Jiangsu, 210012
China
Email: zhangliang1@huawei.com

Qin Wu
Huawei
101 Software Avenue, Yuhua District
Nanjing
Jiangsu, 210012
China
Email: bill.wu@huawei.com