FlowCGAN: Exploratory Study of Class Imbalance for Encrypted Traffic Classification Using CGAN

Pan Wang*, Shuhang Li*, Feng Ye†, Zixuan Wang* and Moxuan Zhang‡
*School of Modern Posts, Nanjing University of Posts & Telecommunications, Nanjing, China
†Department of Electrical & Computer Engineering, University of Dayton, Dayton, OH, USA
‡Schools of International Education, Jinling Institute of Technology, Nanjing, China
Email: *wangpan@njupt.edu.cn, *lish@runtrend.com.cn, †fye001@udayton.edu, *wangzx@runtrend.com.cn
, ‡zhangmoxuan_7@126.com

***Abstract*—With more and more adoption of Deep Learning (DL) in the field of image proessing, computer vision and NLP, researchers have begun to apply DL to tackling with encrypted traffic classification problems. Although these methods can automatically extract traffic features to overcome the difficulty of traditional classification methods like DPI in terms of feature engineering, a large amount of data is needed to learn the characteristics of various types of traffic. Therefore, the performance of classification model always significantly depends on the quality of dataset. Nonetheless, the building of dataset is a time-consuming and costly task, especially encrypted traffic data. Apparently, it is often more difficult to collect a large amount of traffic samples of those unpopular encrypted applications than well-known ones, which leads to the problem of class imbalance between major and minor encrypted application in dataset. In this paper, we proposed traffic data augmentation method called FlowCGAN using Conditional GAN, one of a genre of Generative Adversarial Network (GAN) . As a generative model, FlowCGAN takes the advantage of GAN to generate new traffic samples by learning the characteristics of the traffic data and thereby balancing between major and minor classes of the data set. To verify the feasibility and evaluate the performance of this method, Convolutional Neural Networks(CNN) was designed to classify three datasets: the original unbalanced dataset, the dataset based on Random Over Sampling (ROS) method and the dataset based on FlowCGAN respectively. The experimental evaluation results show that the FlowCGAN method can achieve better performance than the other two in terms of encrypted traffic classification.**

***Index Terms*—encrypted traffic classification, deep augmentation, Conditional Generative Adversarial Network, traffic identification, class imbalance**

I. Introduction

With the rapid development of network technology, the types and quantity of traffic data in cyberspace are increasing. Network traffic identification and classification is a crucial research task in the area of network management and security. It is the footstone of dynamic access control, network resources scheduling, content based billing, intrusion and malware detection etc. High efficient and accurate traffic classification is of great practical significance to provide service quality assurance, dynamic access control and abnormal network behaviors detection. With the widespread adoption of encryption techniques for internet, especially 5G and IoT applications, the growth of portion of encrypted traffic has dramatically posed a huge challenge for QoS, network management and security monitoring. Therefore, studies on encrypted traffic classification not only help to improve the fine-grained network resource allocation based on application, but also enhance security level of network and application.

Traditionally, the evolution of encrypted traffic classification technology has gone through three stages: port matching based, payload matching based and flow statistical characteristics based. Port matching based classification method infers applications' types by assuming that most applications consistently use 'well known' TCP or UDP port numbers, however, the emergence of port camouflage, dynamic port, proprietary protocols with user-defined ports and tunneling technology makes these methods lose efficacy quickly. Payload matching based methods, namely, DPI (Deep Packet Inspection) technology cannot deal with encrypted traffic because of invisible packet content of encrypted traffic, in addition, it incurs high computational overhead and requires manual signatures maintenance [1–3]. As a result, in order to attempt to solve the aboved problems of encrypted traffic identification , flow-based methods emerged, which usually combine statistical or time series traffic features and Machine Learning (ML) algorithms, such as naive bayes(NB), support vector machine(SVM), decision tree, Random Forest(RF), k-nearest neighbor(KNN) [4–7]. Although classical machine learning approach can solve many issues that port and payload based methods cannot solve, it still has some limitations, such as handcrafted traffic features driven by domain-expert, time-consuming, lack of ability of automation, rapidly outdated when compared to the evolution. Unlike most traditional ML algorithms, Deep Learning performs automatic feature extraction without human intervention, which undoubtedly makes it a highly desirable approach for traffic classification, especially encrypted traffic. Recent research work has demonstrated the superiority of DL methods in traffic classification [8], such as MLP [9], CNN [10–14], SAE [15], LSTM [16, 17].

However, due to the different popularity of various applications, the class imbalance problem of traffic samples often occurs when building traffic datasets. That is, the number of popular application samples is much larger than others, which always leads to the misclassifying problems of minor applications and thereby decrease of classifier performance. Imbalanced class distribution of a dataset has posed a serious

challenge to most ML based classifiers which assume a relatively balanced distribution [18]. Network traffic classification is no exception due to the imbalanced property of network traffic data [19, 20], especially encrypted traffic. Therefore, it plays a very crucial role to deal with the problem of imbalanced class distribution of traffic dataset for network traffic classification. However, there are very few studies focusing on traffic data augmentation used for traffic classification to overcome the limitation of class imbalance.

In this paper, we proposed traffic data augmentation method called FlowCGAN using Conditional GAN, one of a genre of GAN. As a generative model, FlowCGAN exploits the benefit of GAN to generate synthesized traffic samples by learning the characteristics of the original traffic data. The synthesized data is then combined with the original (viz. real) data to build the new traffic dataset and thereby keep balance between major and minor classes of the dataset. As a proof of concept, CNN was adopted and designed to classify three encrypted traffic datasets: the original unbalanced dataset, the dataset based on Random Over Sampling (ROS) [21] method and the new augmented dataset based on FlowCGAN respectively. The experimental evaluation results show that classical deep learning based encrypted traffic classifier over our new dataset based on FlowCGAN method can achieve better performance than the other two in terms of encrypted traffic classification.

The rest of this paper is organized as follows. Section II introduces the preliminaries and related works of traffic classification, some current methods for tackling with the problem of imbalanced class data and GAN. Section III illustrates the algorithm of FlowCGAN. Section IV describes the design of the encrypted traffic classification method based on FlowGAN. The experimental results are provided and discussed in Section V. Section VI concludes our work and presents some future works.

REFERENCE

[1] M. Finsterbusch, C. Richter, E. Rocha, J. Muller, and K. Hanssgen, "A survey of payload-based traffic classification approaches," *IEEE Communications Surveys Tutorials*, vol. 16, no. 2, pp. 1135–1156, Second 2014.

[2] P. Wang, F. Ye, and X. Chen, "A smart home gateway platform for data collection and awareness," *IEEE Communications Magazine*, vol. 56, no. 9, pp. 87–93, Sep. 2018.

[3] P. Wang, X. Chen, F. Ye, and Z. Sun, "A smart automated signature extraction scheme for mobile phone number in human-centered smart home systems," *IEEE Access*, vol. 6, pp. 30 483–30 490, 2018.

[4] A. Dainotti, A. Pescape, and K. C. Claffy, "Issues and future directions in traffic classification," *IEEE Network*, vol. 26, no. 1, pp. 35–40, January 2012.

[5] G. Sun, Y. Xue, Y. Dong, D. Wang, and C. Li, "An novel hybrid method for effectively classifying encrypted traffic," in *2010 IEEE Global Telecommunications Conference GLOBECOM 2010*, Dec 2010, pp. 1–5.

[6] P. Velan, M. Čermák, P. Čeleda, and M. Drašar, "A survey of methods for encrypted traffic classification and analysis," *Netw.*, vol. 25, no. 5, pp. 355–374, Sep. 2015. [Online]. Available: http://dx.doi.org/10.1002/nem.1901

[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)*, April 2011, pp. 107–114.

[8] G. Aceto, D. Ciuonzo, A. Montieri, and A. Pescap, "Mobile encrypted traffic classification using deep learning," in *2018 Network Traffic Measurement and Analysis Conference (TMA)*, June 2018, pp. 1–8.

[9] 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. 55 380–55 391, 2018.

[10] M. J. S. M. S. Mohammad Lotfollahi, Ramin Shirali Hossein Zade, "Deep packet: A novel approach for encrypted traffic classification using deep learning," *Available from http://www.arxiv.org*, 2017.

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

[12] and, , and and, "Malware traffic classification using convolutional neural network for representation learning," in *2017 International Conference on Information Networking (ICOIN)*, Jan 2017, pp. 712–717.

[13] 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)*, Dec 2017, pp. 1271–1276.

[14] X. Chen, J. Yu, F. Ye, and P. Wang, "A hierarchical approach to encrypted data packet classification in smart home gateways," in *2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)*, Aug 2018, pp. 41–45.

[15] Z.Wang, "The application of deep learning on traffic identification," *Available from http://www.blackhat.com*, 2015.

[16] 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. 18 042–18 050, 2017.

[17] 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*, vol. 6, pp. 1792–1806, 2018.

[18] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intell. Data Anal.*, vol. 6, no. 5, pp. 429–449, Oct. 2002. [Online]. Available: http://dl.acm.org/citation.cfm?id=1293951.1293954

[19] 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*, ser. SoICT 2017. New York, NY, USA: ACM, 2017, pp. 333–339. [Online]. Available: http://doi.acm.org/10.1145/3155133.3155175

[20] L. Vu, D. Van Tra, and Q. U. Nguyen, "Learning from imbalanced data for encrypted traffic identification problem," in *Proceedings of the Seventh Symposium on Information and Communication Technology*, ser. SoICT '16. New York, NY, USA: ACM, 2016, pp. 147–152. [Online]. Available: http://doi.acm.org/10.1145/3011077.3011132

[21] H. Guo and H. L. Viktor, "Learning from imbalanced data sets with boosting and data generation: The databoost-im approach," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 30–39, Jun. 2004. [Online]. Available: http://doi.acm.org/10.1145/1007730.1007736