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 datasets. Nonetheless, the building of datasets 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 applications in datasets. In this paper, we proposed a novel traffic data augmentation method called FlowCGAN using Conditional GAN, one of a genre of Generative Adversarial Network (GAN) . As a generative model, FlowCGAN exploit the benefit of GAN and CGAN's benefit? to generate new traffic samples by learning the characteristics of the traffic data and thereby balancing between major and minor classes of the datasets. As a proof of concept, Convolutional Neural Network (CNN) was adopted and designed to classify three encrypted traffic datasets: the original unbalanced dataset, the dataset augmented by Random Over Sampling (ROS) method and the new augmented dataset based on FlowCGAN respectively. The experimental evaluation results demonstrate that CNN based encrypted traffic classifier over our new dataset based on FlowCGAN can achieve better performance than the other two in terms of encrypted traffic classification.**

***Index Terms*—encrypted traffic classification, data 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 applications.

Traditionally, the evolution of encrypted traffic classification technology has gone through three stages: port matching, payload matching and flow statistical characteristics based classification methods. Port matching based 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 with 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 approaches 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, DL 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 applications samples is much larger than others, which always leads to the misclassifying problems of minor applications and thereby incurs deterioration 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 is very crucial to address such challenges of imbalanced class distribution of traffic datasets for network traffic classification. However, there are very few studies focusing on traffic data augmentation 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 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.

need to modify: 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.

II. Related works

A. ML and DL based approach of Traffic Classification

Different from port and payload matching methods, ML based classification methods always use payload-independant parameters such as packet length, inter-arrival time and flow duration to circumvent the problems of encrypted content and user's privacy [22]. Many work was carried out using ML algorithms during the last decades. In general, there are two learning strategies used: one is the supervised methods like decision tree, SVM and Naive Bayes, the other is unsupervised approaches like k-means and PCA [23]. Nevertheless, many drawbacks hindered ML based methods widely applied to traffic classification, such as handcrafted traffic features driven by domain-expert, time-consuming, unsuited to 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 [8–17, 24]. The workflow of DL based classfication usually consists of three steps. First, model inputs are defined and designed according to some principles, such as raw packets, PCAP files or flow statistics features. Second, models and algorithms are elaborately chosen according to models' characteristics and aim of the classifier. Finally, the DL classifier is trained to automatically extract the features of traffic.

B. Traditional methods for handling imbalanced data

In general, there are three methods for dealing with class imbalance problem: ***Modifying the objective cost function, Sampling and Generating artificial data*** [20]. The approach of ***modifying objective cost function*** alleviates the problem by means of weighting differently the data samples in minor and major classes, which gives higher score on the minor samples to penalize more intensely on miss-classifiying of the sample in the minor class. Sampling methods include two different ways of ***under-sampling*** and ***over-sampling***, which is to reduce the size of major class by removing some major data samples and raise the ones in the minor class, respectively. Random under sampling (RUS) and Random over sampling (ROS) are two main methods of under-sampling and over-sampling [25]. RUS randomly removes some instances in major class, accordingly, ROS generates some copies of samples of the minor class. However, overfitting problem is always the main drawback of ROS due to generating same copies from the minor class. A classical method for generating artificial data is Synthetic Minority Over-sampling Techinique (SMOTE) in which minority samples are generated by synthetic samples rather than copies [26].

C. The application of GAN and other DL techniques in generating traffic data samples

Due to the great success of GAN applying in images, computer vision and NLP etc., this innovative technique has been already applied to network security recently. A few current studies have shown that GAN has been applied in IDS and Malware detection to generate adversarial attacks to deceive and evade the detection systems [27, 28] and thus effectively improve the performance of malware detection or IDS [27–31]. Correspondingly, as for traffic classification, some researchers have introduced some approaches based on GAN to generate the traffic samples to overcome the imbalaced limitation of network data. In [32], the authors proposed a novel method called auxiliary classifier GAN (AC-GAN) to generate synthesized traffic samples for balancing between the minor and major classes over a well-known traffic dataset NIMS. The AC-GAN took both a random noise and a class label as input in order to generate the samples of the input class label accordingly. The experimental results has shown that their proposed method achieved better performance compared to other methods like SMOTE. However, the NIMS dataset was only composed of SSH and non-SSH two classes. In [33], the authors proposed a novel data augmentation approach based on the use of Long Short Term Memory

(LSTM) network to learn the traffic flow patterns and Kernel Density Estimation (KDE) for replicating the sequence of packets in a flow for classes with less population. The results have shown that this method can improve the performance of DL algorithms over augmented datasets. need to add some description of the drawback of ACGAN and LSTM methods.

III. Generative Adversarial Networks

A. GAN

As an unsupervised learning model, a classic GAN network consists of two parts, the generator G and the discriminator D. The role of the generator is to take random noise as input by learning the characteristic distribution of real data. The discriminator aims at determining whether the data is real or generated by G. The generator G simulates the feature distribution P_g of the real data by the prior distribution $P_z(z)$. The input of the discriminator is the real and generated data, correspondingly, the output $D(x)$ indicates the probability of whether the input data is real or not [34]. During the training process, G and D play a two-player mini-max game until D can't judge whether the sample data is real, which means that the two networks reach the Nash Equilibrium. The objective function of GAN can be expressed by (1):

$$\begin{aligned}\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] \\ +\mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))]\end{aligned} \tag{1}$$

In Equation (1), $P_{data}(x)$ represents distribution of the real data. When training D, the goal is to optimize the probability of TRUE $D(G(z))$ as small as possible and the probability of TRUE $D(x)$ of the real data x as much as possible. When training G, the goal is to make $D(G(z))$ as much as possible. From (1), we can calculate the optimal discriminator as (2). As can be seen from (2) below, when $P_{data}(x) = P_z(z)$, it means that D cannot distinguish whether the sample is true or false, D and G reach the Nash Equilibrium, and the discriminator output is 0.5.

$$D(x) = \frac{P_{data}(x)}{P_{data}(x) + P_z(z)} \tag{2}$$

B. CGAN

In GAN, there is no control over modes of the data to be generated. The conditional GAN changes that by adding the constraint condition y as an additional parameter to the generator G and hopes that the corresponding images are generated. For example, in MNIST, the digit generated by GAN may be either any digit from 0-9 instead of specified one or always output the same digit. Fig. 1 shows the network structure of CGAN.

The principle, structure and training process of CGAN are similar to GAN. The cost function is slightly different as is shown in (3):

$$\begin{aligned}\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x|y)] \\ +\mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z|y)))]\end{aligned} \tag{3}$$

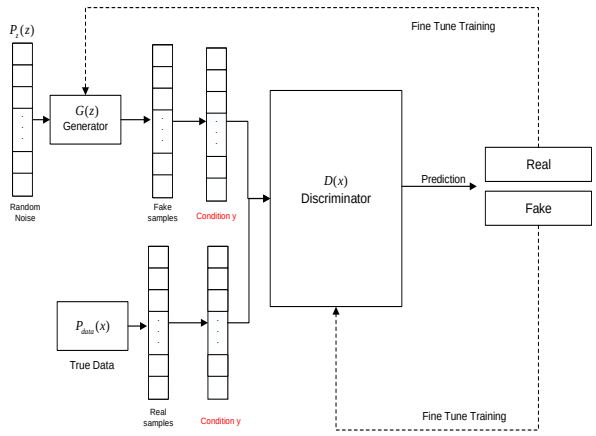

Fig. 1: The network structure of CGAN.

As shown in Fig. 1, CGAN's training process includes the following steps:

- Sampling the real data to obtain $P_{data}(x)$, obtaining the label y corresponding to the sampling data $P_{data}(x)$, feed $P_{data}(x)$ and y into the discriminator D, then updating the parameters according to the output results;
- Generating random noise $P_z(z)$, which is then fed into generator G together with label y in the above step, and G generates synthesized data.
- Feeding the synthesized data and the label y generated in the above step into the discriminator D, and G will optimize the parameters according to the output result of D.
- Repeat the above steps until G and D reach the Nash equilibrium.

C. InfoGAN

In CGAN, the labels y always explicitly are read and passed into the generator $G(z,y)$ and discriminator $D(x,y)$. However, in InfoGAN, the generator and discriminator are $G(z,c)$ and $D(x)$, which c is the latent code representing the semantic features of the datapoints and the noise vector z is the source of noise for the latent variables similar to CGAN. In InfoGAN, c will be learned from x in a DNN instead of initializing it explicitly with labels in CGAN. The codes are then made meaningful by maximizing the Mutual Information between the code and the Generator G output. The objective function of the InfoGAN is as shown in Equation (4).

$$\min_G \max_D V_I(D,G) = V(D,G) - \lambda I(c; G(z,c)) \tag{4}$$

with

$$\begin{aligned}V_{GAN}(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] \\ +\mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z,c)))]\end{aligned} \tag{5}$$

To compute $I(c, G(z, c))$, it is necessary to approximate $p(c|x)$ with a function $Q(c|x)$(Variation Maximization), which is as follows:

$$\begin{aligned} &\mathbb{E}_{x\sim G(z,c),c\sim p(c|x)}[\log p(c|x)] \\ &= \mathbb{E}_{x\sim G(z,c),c\sim p(c|x)}[\log Q(c,x)] + \\ &\mathbb{E}_{x\sim G(z,c),c\sim p(c|x)}[\log \frac{p(c|x)}{Q(c,x)}] \\ &= \mathbb{E}_{x\sim G(z,c),c\sim p(c|x)}[\log Q(c,x)] + \\ &\mathbb{E}_{x\sim G(z,c),c\sim p(c|x)}[D_{KL}(p(c|x))||Q(c,x)] \\ &\geq \mathbb{E}_{x\sim G(z,c),c\sim p(c|x)}[\log Q(c,x)] \end{aligned} \tag{6}$$

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

[22] D. C. Sicker, P. Ohm, and D. Grunwald, "Legal issues surrounding monitoring during network research," in *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, ser. IMC '07. New York, NY, USA: ACM, 2007, pp. 141–148. [Online]. Available: http://doi.acm.org/10.1145/1298306.1298307

[23] T. T. T. Nguyen and G. Armitage, "A survey of techniques for internet traffic classification using machine learning," *IEEE Communications Surveys Tutorials*, vol. 10, no. 4, pp. 56–76, Fourth 2008.

[24] D. Li, Y. Zhu, and W. Lin, "Traffic identification of mobile apps based on variational autoencoder network," in *2017 13th International Conference on Computational Intelligence and Security (CIS)*, Dec 2017, pp. 287–291.

[25] N. Japkowicz, "Learning from imbalanced data sets: A comparison of various strategies." AAAI Press, 2000, pp. 10–15.

[26] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *J. Artif. Int. Res.*, vol. 16, no. 1, pp. 321–357, Jun. 2002. [Online]. Available: http://dl.acm.org/citation.cfm?id=1622407.1622416

[27] Z. Lin, Y. Shi, and Z. Xue, "Idsgan: Generative adversarial networks for attack generation against intrusion detection," *CoRR*, vol. abs/1809.02077, 2018.

[28] W. Hu and Y. Tan, "Generating adversarial malware examples for black-box attacks based on gan," *CoRR*, vol. abs/1702.05983, 2017.

[29] J.-Y. Kim, S.-J. Bu, and S.-B. Cho, "Malware detection using deep transferred generative adversarial networks," 10 2017, pp. 556–564.

[30] M. Salem, S. Taheri, and J. S. Yuan, "Anomaly generation using generative adversarial networks in host based intrusion detection," *CoRR*, vol. abs/1812.04697, 2018.

[31] M. Rigaki and S. Garcia, "Bringing a gan to a knife-fight: Adapting malware communication to avoid detection," in *2018 IEEE Security and Privacy Workshops (SPW)*, May 2018, pp. 70–75.

[32] L. Vu, C. Thanh Bui, and U. Nguyen, "A deep learning based method for handling imbalanced problem in network traffic classification," 12 2017, pp. 333–339.

[33] R. Hasibi, M. Shokri, and M. Dehghan, "Augmentation scheme for dealing with imbalanced network traffic classification using deep learning," *ArXiv*, vol. abs/1901.00204, 2019.

[34] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Networks," *arXiv e-prints*, p. arXiv:1406.2661, Jun 2014.

[35] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," *arXiv e-prints*, p. arXiv:1411.1784, Nov 2014.