ISSN: 2661-3719(Print); 2661-3727(Online)

基于对抗训练的加密恶意流量检测技术研究

周　悦
上海市计算机软件测评重点实验室　上海计算机软件技术开发中心　201112

摘　要：针对加密恶意流量的检测问题，本文研究了基于对抗训练的加密恶意流量检测技术。首先使用真实的样本训练基于深度学习技术的原始检测模型，随后根据真实样本生成对抗样本，并使用对抗样本继续训练模型。实验表明本文所提方法能够有效减少数据集对深度学习模型的影响，增强检测模型对加密恶意流量的检测能力。
关键词：加密流量；网络入侵检测；深度学习；对抗训练

Research on encryption malicious traffic detection technology based on adversarial training

Yue Zhou
Shanghai Key Laboratory of Computer Software Evaluation Shanghai Computer Software Technology Development Center 201112

Abstract: This paper addresses the detection problem of encrypted malicious traffic and investigates the use of adversarial training for encrypted malicious traffic detection. Initially, real samples are employed to train the initial detection model based on deep learning techniques. Subsequently, adversarial samples are generated based on the real samples, and the model is further trained using these adversarial samples. The experiments demonstrate that the proposed method effectively reduces the impact of the dataset on the deep learning model and enhances the detection capability of the model for encrypted malicious traffic.
Keywords: encrypted traffic; Network intrusion detection; Deep learning; Confrontation training

前言：

流量加密技术是一种重要的信息安全手段，然而流量加密不仅保护了正常的互联网用户，也为恶意的网络攻击者提供了掩护[1]。恶意流量经过加密不仅隐藏了它传输的数据内容，也改变了部分与流量大小相关的统计特征，使得大量传统的入侵检测手段对其失效。

传统的入侵检测主要通过匹配特定字符串等方法来识别攻击流量[2]，但是这种方法只能识别已知的攻击，并且由于加密流量隐藏了传输的数据，使得这种方法对加密流量基本失效[3]。随后，研究人员针对基于统计特征的检测方法进行了研究，该方法可以无视流量的数据内容，仅提取与流量的数据大小和传输时间相关的各种特征来进行检测。这种方法对加密流量具有一定的有效性，但是加密对数据的改变会在一定程度上影响与流量大小相关的统计特征，并且特征的设计对检测结果的影响非常大[4]。随着深度学习技术在各个领域崭露头角，该技术同样进入了入侵检测研究人员的视野[5]。研究表明该方法能够有效对加密流量进行检测，并且不再需要人工设计特征。但是该类方法的效果极大受限于数据集的数量和质量。

针对加密恶意流量检测的问题，本文提出了基于对抗训练的加密恶意流量检测技术研究。将流量数据转化为图片，训练深度学习分类器。随后针对分类器构造对抗样本，使用对抗样本和真实样本的混和数据继续训练分类器。实验结果证明，基于同样的数据集进行训练，本文的方法能够成功提高深度学习分类器的检测性能。

一、基于对抗训练的加密恶意流量检测方法

本文提出的方法主要分为两部分，第一部分是基于深度学习模型，使用真实的流量数据训练原始检测模型；第二部分是针对原始检测模型，在真实流量数据的基础上生成对抗样本，用对抗样本和真实数据的混合数据集继续训练原始检测模型，获得最终的加密恶意流量检测模型。

1. 原始检测模型生成

原始检测模型的生成步骤如图 1 所示。

ISSN: 2661-3719(Print); 2661-3727(Online)

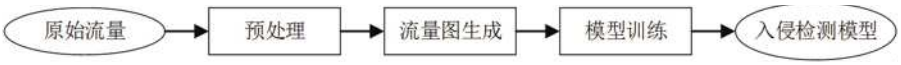
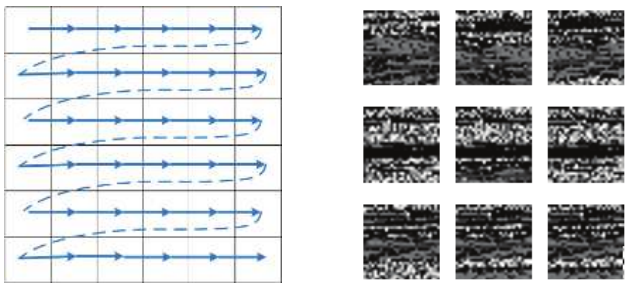


图1　基于深度学习的加密恶意流量检测流程

预处理：预处理首先将庞大流量抓包文件按照会话进行切分，一个会话表示通信双发一次完整的通信，因此本文使用会话作为入侵检测的基本单元。依据通信双方的IP地址、端口号和传输层协议来识别属于同一会话的流量包。切分完成后，通过编程去除掉IP地址和端口号这些可能使得深度学习模型过拟合的信息，并且去除了SSL/TLS加密流量协议中的SNI（Server Name Indication）字段，该字段并非属于恶意流量的特征，同样可能引起模型的过拟合。

流量图生成：流量是以2位16进制数的形式进行传播的。本方法将流量会话中的每个2位16进制数转化成10进制数作为像素的灰度，然后按照图2（a）的方式排列数据，排列时只取会话的前1024个字节[7]，就可以获得图2（b）所示的流量图，一张流量图就是深度学习模型的一个输入。图2（b）总共展示了9张转化后的流量图。



（a）流量图构造方法　　（b）流量图示意

图2　流量图的构造方法与流量图示意

模型训练：将正常流量和恶意流量的流量图混合，打好标签训练深度学习模型，从而获得原始入侵检测模型，用来判断流量的恶意与非恶意。本文的实验使用了CaForest[8]和GoogLeNet[9]两种深度学习模型进行实验。CaForest模型基于森林模型的组合和堆叠构成，是一种有别于神经网络的特殊结构；GoogLeNet是一种经典的CNN模型，尽管它在性能上比于当前主流的CNN模型较弱，但是与当前主流的CNN模型相比体量更小，速度更快，更加适合对实时性要求较高的入侵检测任务。

2.基于对抗训练的检测模型加强

生成原始检测模型后，通过对抗模型对原始模型进行加强，流程图如图3所示。

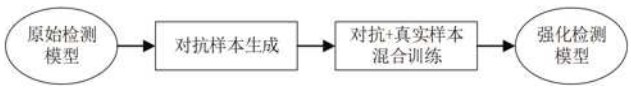


图3　基于对抗训练的加密恶意流量检测模型强化流程

对抗样本生成：这一步主要是将生成的恶意流量图转化对抗样本，使得原始模型将对抗样本判断为正常流量。实验种采用了改进后的One Pixel Attack（OPA）算法[10]生成对抗样本。OPA算法的优点是可以只对原始图片的有限个像素点（最少为1个）进行改变，能够最大程度的保留样本原始数据。OPA算法的核心是通过遗传进化算法来搜寻要修改的像素点和修改后的像素值，本文对OPA的改进主要有两点：（1）在每一轮中只取最优的个体进行变异，这一步是为了加快算法的收敛速度；（2）遗传每进行3轮，用随机生成的个体对种群中10%的个体进行替换，这一步是为了防止搜寻陷入局部最优。

对抗+真实样本混合训练：这一步中将原始模型作为预训练模型，使用原始流量图和生成的对抗流量图的混合数据集对模型继续进行训练，最后得到强化后的检测模型。

二、实验结果

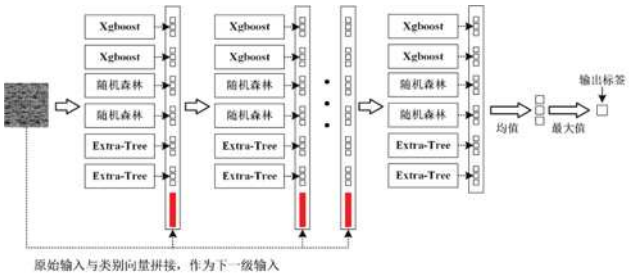
1.实验配置

本文的实验数据来自MCFP dataset[11]和ISCX VPN-nonVPN dataset[12]两种公开数据集，其中ISCX VPN-nonVPN dataset中为正常的加密流量，MCFP dataset为恶意的加密流量。数据集经处理后获得的样本数量如表1所示。

表1　样本数量分布

类别	正常	恶意
数量	3020	1000

实验使用的CaForest模型结构如图4所示。模型的每一级由3种共6个不同的森林模型组成。GoogLeNet模型则采用了论文[9]中的结构。



2.评价指标

本文采用了以下三项指标来评价检测模型的性能：

$$ACC=\frac{TP+TN}{TP+FP+TN+FN}\times 100\% \quad (1)$$

$$DR=\frac{TP}{TP+FN}\times 100\% \quad (2)$$

$$FAR=\frac{FP}{FP+TN}\times 100\% \quad (3)$$

其中TP、FP、TN、FN分别表示真阳性、假阳性、真阴性、假阴性。ACC表示模型分类的整体准确率；DR表示模型对恶意样本的检测成功率；FAR表示模型检测出的恶意样本中，误报的比率。

3.OPA算法改进实验

本文的方法对原始的OPA算法进行了改进，本实验用来验证改进的有效性。参考文献[12]，使用了成功率和平均欺骗度两个指标来评价算法的性能。计算方法如公式（4）（5）所示。

成功率=对抗成功样本数量/原始样本数量 ×100%

平均欺骗度=欺骗度总和/对抗成功样本数量 ×100%

对抗生成是通过对原始恶意样本进行扰动，使得扰动后得到的对抗样本被原始检测模型检测为非恶意样本。因此若模型认为一个对抗样本为非恶意流量的概率超过50%，则该对抗样本视为一个对抗成功样本。欺骗度为模型输出的一个样本为非恶意流量概率，例如一个恶意样本经过对抗生成后，模型认为其有80%概率是正常样本，则该对抗样本的欺骗度为0.8。

本实验对数据集中所有的原始恶意样本进行了对抗生成，统计数据得到的实验结果如表2所示。由实验结果可知，改进后的OPA性能得到了极大提升，在成功率和平均欺骗度分别比原始的OPA算法高出了43%和13%，并且算法耗时在实验数据集上减少了854秒。

表2　对抗样本生成算法性能对比

	成功率（%）	平均欺骗度（%）	时间（s）
OPA	42.5	74	2304
改进OPA	85.5	87	1450

4.模型强化实验

模型强化的实验结果如表3所示，由表3可知，检测模型经过强化后，性能大大提升，其中基于CNN的检测模型DR增加了13.33%之多，而FAR仅仅增加了1.71%。而基于CaForest的检测模型三项数据均有变优。

表3　不同模型强化前后加密恶意流量检测能力对比

模型	是否强化	ACC	DR	FAR
GoogLeNet	强化前	88.53%	71.31%	5.79%
	强化后	90.50%	84.64%	7.50%
CaForest	强化前	99.21%	98.00%	0.40%
	强化后	99.51%	98.80%	0.27%

三、结论

本文提出了基于对抗训练的加密恶意流量检测方法。首先基于深度学习模型构建针对加密流量的分类模型，随后通过改进后的OPA算法生成对抗样本，并通过对抗训练强化分类模型。在真实数据集上的实验证明了本文对OPA算法的改进十分有效。实验基于两种不同的深度学习模型构建加密恶意流量检测模型，实验结果证明了本文提出的方法能够有效提高模型对加密恶意流量的检测能力。

参考文献：

[1] Threats in encrypted traffic [EB/OL]. https://blogs.cisco.com/security/threats-in-encrypted-traffic.

[2] S. Lee, J. Park, S. Yoon and M. Kim. High performance payload signature-based Internet traffic classification system[C]. 2015 17th Asia-Pacific Network Operations and Management Symposium (APNOMS), 2015: 491-494.

[3] S. Rezaei and X. Liu. Deep Learning for Encrypted Traffic Classification: An Overview[J]. in IEEE Communications Magazine, 2019, 57(5): 76-81.

[4] R. Sommer and V. Paxson. Outside the Closed World: On Using Machine Learning for Network Intrusion Detection[C]. 2010 IEEE Symposium on Security and Privacy, 2010: 305-316.

[5] J. A. Abraham and V. R. Bindu. Intrusion Detection and Prevention in Networks Using Machine Learning and Deep Learning Approaches: A Review[C]. 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), Coimbatore, India, 2021: 1-4

[6] J. Lansky et al. Deep Learning-Based Intrusion Detection Systems: A Systematic Review [J]. IEEE Access, 2021, 9:101574-101599.

[7] L. Vu, et al. Time Series Analysis for Encrypted Traffic Classification: A Deep Learning Approach[C]. 2018 18th International Symposium on Communications and Information Technologies (ISCIT), 2018: 121-126.

[8] Zhou, Zhi-Hua, and Ji Feng. Deep forest: Towards an alternative to deep neural networks [EB/OL]. 2017, arXiv preprint arXiv:1702.08835.

[9] C. Szegedy, V. Vanhoucke, S. Ioffe, et.al. Rethinking the Inception Architecture for Computer Vision [C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, 2818-2826.

[10] J. Su, D. V. Vargas and K. Sakurai. One Pixel Attack for Fooling Deep Neural Networks[J]. in IEEE Transactions on Evolutionary Computation, 2019, 23(5): 828-841.

[11] Maria Jose Erquiaga, MCFP dataset[OL], https://mcfp.felk.cvut.cz/publicDatasets/

[12] ISCX Vpn-nonVpn dataset[OL], http://www.unb.ca/cic/datasets/vpn.html