基于对抗样本的网络欺骗流量生成方法

胡永进[1]，郭渊博[1]，马骏[1]，张晗[1,2]，毛秀青[1]
（1. 信息工程大学密码工程学院，河南 郑州 450001；2. 郑州大学软件学院，河南 郑州 450000）

摘 要：为了应对流量分类攻击，从防御者的角度出发，提出了一种基于对抗样本的网络欺骗流量生成方法。通过在正常的网络流量中增加扰动，形成欺骗流量的对抗样本，使攻击者在实施以深度学习模型为基础的流量分类攻击时出现分类错误，欺骗攻击者从而导致攻击失败，并造成攻击者时间和精力的消耗。采用几种不同的扰动生成方法形成网络流量对抗样本，选择 LeNet-5 深度卷积神经网络作为攻击者使用的流量分类模型实施欺骗，通过实验验证了所提方法的有效性，为流量混淆和欺骗提供了新的方法。

关键词：对抗样本；网络流量分类；网络欺骗；网络流量混淆；深度学习
中图分类号：TP393
文献标识码：A
doi: 10.11959/j.issn.1000−436x.2020166

Method to generate cyber deception traffic based on adversarial sample

HU Yongjin[1], GUO Yuanbo[1], MA Jun[1], ZHANG Han[1,2], MAO Xiuqing[1]
1. Department of Cryptogram Engineering, Information Engineering University, Zhengzhou 450001, China
2. Software College, Zhengzhou University, Zhengzhou 450000, China

Abstract: In order to prevent attacker traffic classification attacks, a method for generating deception traffic based on adversarial samples from the perspective of the defender was proposed. By adding perturbation to the normal network traffic, an adversarial sample of deception traffic was formed, so that an attacker could make a misclassification when implementing a traffic analysis attack based on a deep learning model, achieving deception effect by causing the attacker to consume time and energy. Several different methods for crafting perturbation were used to generate adversarial samples of deception traffic, and the LeNet-5 deep convolutional neural network was selected as a traffic classification model for attackers to deceive. The effectiveness of the proposed method is verified by experiments, which provides a new method for network traffic obfuscation and deception.

Key words: adversarial sample, network traffic classification, cyber deception, network traffic obfuscation, deep learning

1 引言

网络流量分类技术作为增强网络可控性的基础技术之一，在帮助研究人员了解流量分布、优化网络传输、提高网络服务质量的同时，也常被攻击者用于对目标网络流量进行监控，攻击者通过对网络流量进行分类确定其所属的应用类型（如邮件类、多媒体类、网站类等），进而根据分类结果实施流量拦截并可进一步实施网站指纹攻击。尤其是随着机器学习和深度学习被用于网络流量分类领域[1]，基于它们的网络流量分类技术为攻击者提供了更便利的条件，往往能获得极高的分类准确率。一种典型的攻击者使用基于深度学习的网络流量分类方法实施攻击的场景如图 1 所示。虽然深度学

收稿日期：2020−03−20；**修回日期**：2020−06−16
通信作者：张晗，zhang_han@zzu.edu.cn
基金项目：信息保障技术重点实验室开放基金资助项目（No.KJ-15-108）
Foundation Item: Foundation of Science and Technology on Information Assurance Laboratory (No.KJ-15-108)

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

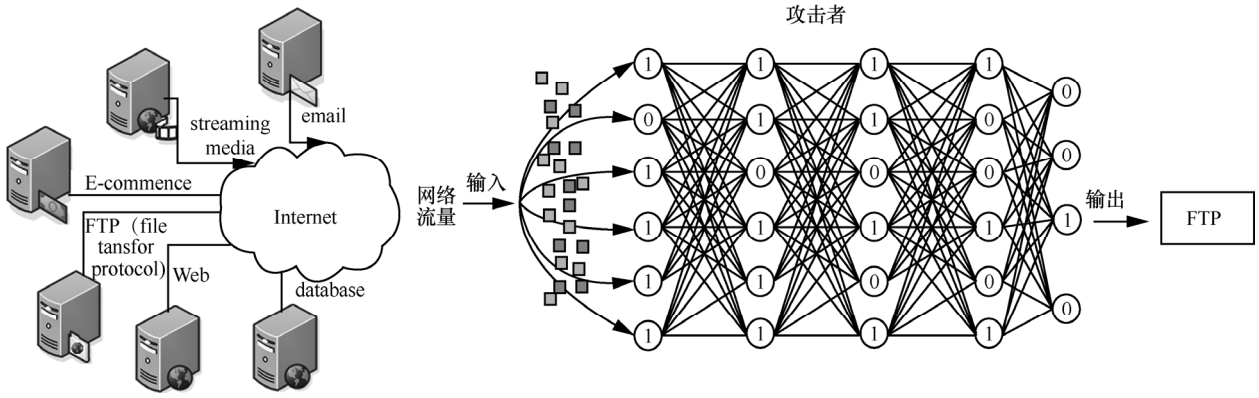


图 1　攻击者使用基于深度学习的网络流量分类方法实施攻击

习在流量分类领域的应用可以提高分类的准确率，并且已经在图像识别、自然语言处理等多个领域展现了巨大的潜力，但随着 Szegedy 等[2]在计算机视觉领域中提出了对抗样本这一概念，对抗包括卷积神经网络（CNN, convolutional neural network）在内的深度学习模型引起了学者们的兴趣。

Szegedy 等[2]在研究图像识别时发现，当对学习模型的输入样本添加一些人眼无法察觉的细微干扰时，会导致模型以高置信度给出一个错误的输出。针对深度学习模型，在原数据集中添加细微扰动所形成的样本被称为对抗样本。从攻击的角度，对抗样本在计算机视觉领域应用最直接，例如在人脸识别、自动驾驶等方面，攻击者可以通过添加人眼无法识别的扰动引发图像识别的分类错误，造成人脸识别或交通标志识别失败[3]；在信息安全领域，欺骗基于神经网络的恶意软件检测模型，从而逃避检测[4]。从防御的角度，对抗样本也具有非常高的价值，主要表现在以下两点：第一点是可以针对深度学习模型提前生成对抗训练样本，提高模型的稳健性以应对可能的对抗样本攻击[5]；第二点是可以针对攻击者使用的深度学习模型，利用对抗样本欺骗其攻击模型，造成攻击的不确定性，并提高攻击成本，使攻击者延缓或取消攻击。本文从对抗样本防御价值的第二点出发，针对攻击者的网络流量分类攻击，通过增加扰动生成网络流量的对抗样本，从而形成网络欺骗流量，引发攻击者的流量分类错误。

本文应用对抗样本的概念应对攻击者发起的网络流量分类攻击，采用 3 种典型的扰动生成方法形成欺骗流量对抗样本，对攻击者采用的基于深度学习的流量分类攻击模型进行欺骗，造成其分类错误，从而导致攻击失败。首先，将对抗样本的概念应用于主动防御中的网络欺骗流量研究，对比不同的对抗样本的欺骗效果；其次，与攻击者通过对抗样本实施攻击相反，本文以对抗样本为防御手段欺骗攻击者的分类模型，可以看作主动防御中的“防中有攻”；最后，针对 LeNet-5 深度卷积神经网络分类模型，即在攻击者分类模型已知（白盒）的条件下进行实验，验证对抗样本在流量分类中的效果。

2　相关工作

2.1　流量分类

根据网络流量的颗粒度，对网络流量分类的研究主要针对表 1 所示的 3 个层面[6]。

表 1　网络流量分类中的研究对象

流量颗粒度	关注点
packet 级	主要关注数据分组（packet）的特征及其到达过程，如数据分组大小分布、数据分组到达时间间隔的分布等
flow 级	主要关注流（flow）的特征及其到达过程，可以为一个 TCP（transmission control protocol）连接或者一个 UDP（user datagram protocol）流
stream 级	主要关注主机对及其之间的应用流量，通常指一个由源 IP 地址、目的 IP 地址、应用协议组成的三元组，适用于更粗粒度上研究骨干网的长期流量统计特性

在上述 3 个层面的流量分类中，flow 级的网络流量是流量分类使用最广泛的对象，其基本思想如图 2 所示。本文以 flow 级的流量作为原始数据，通过生成对抗流量样本对攻击者的流量分类进行欺骗，欺骗的目标是使用基于深度学习的分类模型的攻击者。基于深度学习的分类方法假设对于某些类别的应用，其网络层的统计特征（如流持续时间的

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

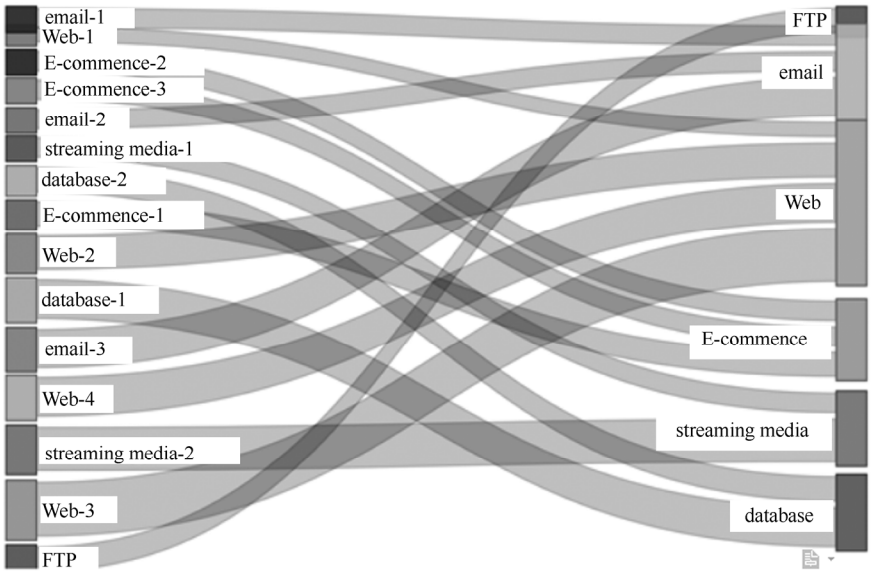


图 2　flow 级的网络流量分类

分布）是唯一的，使用决策树、朴素贝叶斯、支持向量机、关联规则学习、神经网络、遗传算法等方法构造分类模型，从而进行分类，具有使用场景广、分类准确率高、能够分类加密数据流量等特点。

基于机器学习的流量分类的主要思想是构建流量的统计属性联合作为指纹进行分类。文献[7]将机器学习应用于流量分类，认为流的字节可以看作图片中的像素，使用在图像识别领域表现非常好的深度学习方法进行流量分类。文献[8]融合特征提取、特征选择和分类为一个端到端的框架，采用一阶卷积神经网络对不同行为的负载字节进行计算，构建指纹。文献[9]利用匿名化网络特性，采用长度序列和方向序列作为深度学习网络，如 SAE（stacked auto-encoder）、CNN、长短期记忆（LSTM, long short-term memory）网络的输入，从而分类访问网页。文献[10]将表征学习的方法应用于恶意流量分类领域，将原始流量数据视为图片，然后使用擅长图片分类任务的卷积神经网络执行分类，最终达到恶意流量分类的目的。这些研究证明了深度学习在流量分类上的可行性，同时为对抗基于深度学习的流量分类提供了方向。

在对抗或欺骗上述基于深度学习的流量分类的研究上，文献[11]提出了加载背景流量的防御方法，并在 TOR（the onion router）和 JAP（Java anonymous proxy）上进行验证；文献[12]验证了加密协议字节填充的加密流量分类对抗的效果；文献[13]使用不同的真实流量作为访问网站时的噪声；文献[14]提出了 walkie-talkie 以单工方式加载网站来混淆 burst 特征。上述研究主要通过修改流量的通信特征实现对抗，多聚焦于如何逃逸以避免被攻击者发现，伪装和欺骗的能力有限，对抗性不足。目前与本文最接近的研究是隐私保护领域的流量伪装与混淆。其中，obfsproxy（obfuscated proxy）模糊代理软件[15]使 SSL（secure socket layer）或 TLS（transport layer security）加密流量看起来像未加密的 HTTP 或即时通信流量；文献[16]发布了 TOR 的传输层插件 SkypeMorph，将 TOR 客户端与网桥的通信流量填充到 Skype 视频通话流量中，用于对抗流量统计分析；文献[17]提出了黑盒分析流量分类规则，通过多次测试推断流量分析识别规则，从而修改通信数据分组进行逃逸。但这些研究均未将对抗样本的概念应用到对抗流量分析的研究中，也没有将其作为防御者的一种欺骗防御方法进行讨论，本文重点研究了这一问题。

2.2　对抗样本

生成对抗样本的关键在于计算和生成对抗扰动。在计算机视觉领域，扰动需要满足添加到原始图像后人眼不可见并可迷惑原有分类模型的要求。本文中，针对流量分类模型的欺骗，虽然生成的扰动不必满足“人眼不可见”的条件，但仍需要达到一定的条件（如带宽要求等）。

多数工作主要研究了引发图像分类模型的对抗扰动[18]。Szegedy 等[2]发现了深度神经网络在图像分类领域存在的弱点，提出了对抗样本的概念，

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

并将对抗性扰动生成描述为优化问题。Goodfellow 等[19]提出了一种最优的最大范数约束扰动的方法，称为快速梯度符号法（FGSM, fast gradient sign method），以提高计算效率，并证明了神经网络高维度线性是导致对抗样本较好的根本原因。Kurakin 等[20]提出了一种基本迭代方法，使用 FGSM 迭代地产生扰动。Moosavi-dezfooli 等[21]发现图像分类模型中存在与特定图像无关的对抗性扰动，即通用扰动，这种通用扰动可以使分类模型对所有增加了该扰动的图片分类错误。Athalye 等[22]发现 3D 打印的真实世界中的物体也可以欺骗深度网络分类器。Moosavi-dezfooli 等[23]提出的 DeepFool 进一步提高了对抗性扰动的有效性。类似于文献[21]，Metzen 等[24]为语义分割任务提出了通用对抗扰动（UAP, universal adversarial perturbation），扩展了 Kurakin 等[20]的迭代 FGSM 攻击，更改了每个像素预测的标签。Mopuri 等[25]寻求数据独立的通用扰动，不从数据分布中采样，提出了一个新的无数据目标的算法来生成通用对抗扰动，称为 FFF（fast feature fool）算法。文献[26]提出的 GDUAP（generalizable data-free universal adversarial perturbations）算法改善了攻击的效果，让多个不同架构和参数分类模型产生错误分类，并证明了 GDUAP 算法在跨计算机视觉任务上的有效性。目前，除了研究对计算机视觉中的分类、识别任务的攻击之外，研究者还在研究对其他领域和方向的攻击，但未发现对网络流量分类方向实施攻击的研究。

3 安全模型

3.1 攻击者模型

本文假设攻击者能够观测到 2 个主机节点之间的 flow 级别的流量，并能够提取数据分组大小、内部数据分组到达时间等流量的特征，通过使用这些流量对分类模型进行训练以推断防御者使用的应用类型，从而进行流量分类。根据文献[27]，建立攻击者模型，具体描述如下。攻击者企图将观测到的流量 TF 分类为应用类型集合 C 中的第 i 类，$C=\{c_1,c_2,c_3,\cdots,c_i,\cdots,c_n\}$；流量 TF 的特征集合为 X，$X=\{x_1,x_2,x_3,\cdots,x_m\}$；分类模型的分类函数为 $F(x_i)$，输出值为属于应用集合 C 中第 i 类的概率。相关符号定义与说明如表 2 所示。

3.2 防御模型

防御者依据流量 TF 通过生成扰动 P 形成欺骗流量 TD。本文使用不同的生成方法计算不同的扰动，形成不同的欺骗流量 TD，从中提取特征集合 $X'=\{x'_1,x'_2,x'_3,\cdots,x'_m\}$，使攻击者的分类函数的输出 $F(x'_i)$ 与原输出 $F(x_i)$ 不同，即攻击者错误地将流量由第 i 类分类为第 i' 类。相关符号定义与说明如表 3 所示。

表 2 攻击者模型相关符号定义与说明

符号	描述	说明
TF	攻击者观测到的流量	—
X	流量 TF 的特征集合	$X=\{x_1,x_2,x_3,\cdots,x_m\}$
C	流量 TF 对应的应用类型集合	$C=\{c_1,c_2,c_3\cdots,c_i,\cdots,c_n\}$
$F(x_i)$	分类模型的分类函数	输入值为流量 TF，输出值为流量 TF 属于应用类型集合 C 中第 i 个应用类型的概率

表 3 防御者模型相关符号定义与说明

符号	描述	说明
P	扰动	—
TD	防御者在流量 TF 内加入扰动 P 后形成的对抗样本，称为欺骗流量	—
X'	欺骗流量 TD 的特征集合	$X'=\{x'_1,x'_2,x'_3,\cdots,x'_m\}$
$F(x')$	分类模型的分类函数	输入值为欺骗流量 TD，输出值为欺骗流量 TD 属于应用类型集合 C 中第 i' 个应用类型的概率

3.3 扰动的生成方法

文献[28]将扰动生成方法分为全像素添加扰动和部分像素添加扰动，并在此基础上又分为目标针对性和非目标针对性、黑盒测试和白盒测试、肉眼可见和肉眼不可见。结合流量分类的特点，本文采用的扰动生成方法为全像素添加扰动，在已知攻击者使用分类器的参数和内部结构的情况下生成流量对抗样本，欺骗攻击者的分类器产生错误分类，不需要产生指定的错误分类类别。基于上述特点，本文使用的 3 种扰动生成方法如下。

1) FGSM

FGSM 由 Goodfellow 等[19]提出，是生成对抗样本的基本方法之一，其基于梯度下降原理通过在梯度方向上添加增量来诱导网络对生成的图片进行错误分类。扰动计算式为

$$P=\varepsilon\text{sign}(\nabla\Im(\theta,x,y)) \tag{1}$$

其中，ε 为限制扰动过大的约束条件，$\|P\|<\varepsilon$（不

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

同的范数效果不同）；$\nabla\Im(\theta,x,y)$ 为训练神经网络时的损失函数，θ 为分类模型的参数，x 和 y 分别为模型的输入和输入对应的正确标签。$\text{sign}(\nabla\Im(\theta,x,y))$ 为点 x 处损失函数的梯度方向，ε 可看作在该方向上的偏移量级。

2) DeepFool 方法

DeepFool 方法是 Moosavi-dezfooli 等[23]提出的一种非针对目标的对抗样本生成方法，对深度网络有很强的对抗性和稳健性。其通过迭代计算生成最小规范对抗扰动，将位于分类边界内的图像逐步推到边界外，直到出现错误分类。该方法生成的扰动比 FGSM 更小。假设分类器的分类函数为 $f(x)=\boldsymbol{w}^{\mathrm{T}}x+b$，$\boldsymbol{w}$ 为权重，b 为截距，可知其仿射平面为 $\Gamma=\{x:\boldsymbol{w}^{\mathrm{T}}x+b=0\}$。当在某一点 x_0 加入扰动 ρ 形成的向量垂直于平面 Γ，则加入的扰动最小（为 ρ_*）且符合迭代要求，如式(2)所示。

$$\rho_*(x_0):=\arg\min\|\rho\|_2$$
$$\text{s.t.}\,\text{sign}(f(x_0+\rho))\neq\text{sign}(f(x_0))=-\frac{f(x_0)}{\|\boldsymbol{w}\|_2^2} \tag{2}$$

3) C&W 方法

Carlini 和 Wagner[29]基于 FGSM、L-BFGS（limited-memory Broyden-Fletcher-Goldfarb Shanno）和 JSMA（Jacobian-based saliency map approach）这 3 种方法提出了 C&W（Carlini and Wagner）方法，其在范数 l_0、l_2、l_∞ 上均有较大改善。以 l_2 范数为例，C&W 方法如式(3)所示。

$$\min\left\|\frac{1}{2}\left((\tanh(w)+1)-x\right)\right\|_2^2+\text{cf}\left(\frac{1}{2}\left(\tanh(w)+1\right)\right)$$
$$\text{where}\,f(x')=\max(\max\{\boldsymbol{Z}(x')_i:i\neq t\}-\boldsymbol{Z}(x')_i-k) \tag{3}$$

其中，扰动 P 由 $\frac{1}{2}\left(\left(\tanh(w)+1\right)-x\right)$ 计算得到，$\tanh()$ 函数利用优化将对抗样本映射到 $[-\infty,+\infty]$；$Z(x)$ 表示样本 x 通过模型未经过 softmax 函数的输出向量，$Z(x')_i$ 表示欲错误形成的类别 t 的逻辑表示；$\max\{Z(x')_i:i\neq t\}$ 表示分类 $i\neq t$ 的逻辑表示，$\max\{Z(x')_i:i\neq t\}-Z(x')_i$ 表示分类错误最小；k 表示可调节的超参数，控制生成的对抗样本的置信度。

式(3)设置了一个特殊的损失函数来衡量输出误差，其由两部分组成，$\frac{1}{2}((\tanh(w)+1)-x)$ 保证所得扰动最小，$\text{cf}\left(\frac{1}{2}(\tanh(w)+1)\right)$ 保证生成的对抗样本可使模型分类错误的那一类的概率更高。由于可以调节生成样本的置信度，C&W 方法可以生成强对抗样本，增强了其对抗迁移性，具备实现黑盒攻击的能力。

3.4　欺骗流量对抗样本生成算法

基于上述分析与扰动生成算法，本文设计了欺骗流量对抗样本生成算法，如算法 1 所示。

算法 1　欺骗流量对抗样本生成算法

输入　正常网络流量 TF

输出　欺骗流量对抗样本 TD

begin

1) Preprocess(TF) //预处理 TF

2) {

3) Extract Features (TF)//提取特征 X

4) TransArfftoIDX(TF)//将 TF 由 arff 格式转化为 IDX 格式

5) Normalized(TF)//消除各特征量纲，归一化到 [0, 255]

6) }

7) Reshape(TF)//将多类特征的每一类特征值作为一个灰度值

8) Visualization(TF)//添加 0，扩充至 256，形成 16×16 的矩阵，将流量可视化

9) Training(TF, mode)//构建和训练模型

10) Test(TF)//测试正常网络流量的准确率

11) CraftingPerturbation(method)//调用不同的扰动生成方法，生成扰动

12) TD=GenerateAdvSample()// TD = TF + P，将扰动和原始流量进行叠加，生成欺骗流量 TD

13) Visualization(TD)//将 TD 与步骤 5)的结果进行比较

14) result=Evaluate(TD)//评估生成的欺骗流量 TD

15) if (result) repeat 9)～14)//TD 不满足要求，重复执行步骤 9)～步骤 14)

16) else return TD //输出欺骗流量对抗样本

end

算法 1 流程如图 3 所示。

算法 1 中首先需要对真实流量 TF 进行预处理和归一化处理。预处理 Preprocess(TF)利用不同的数据集对真实流量 TF 进行清洗，并提取 TF 的流量特征；参考图像识别领域中对图像数据的读取格式，将流量数据集的数据格式转化为图像数据格式，通

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

过 TransArfftoIDX(TF)将流量数据集的数据格式由 arff 转换为 IDX 格式；通过 Normalize(TF)对数据集所有流量的每一项特征进行数值归一化，消除流量数据集各个特征之间的量纲关系，使不同特征之间具有可比性，这样就可以使用 Reshape(TF)和 Visualization(TF)将真实流量的各类特征值视为图像的灰度值，进而以图像的方式可视化。

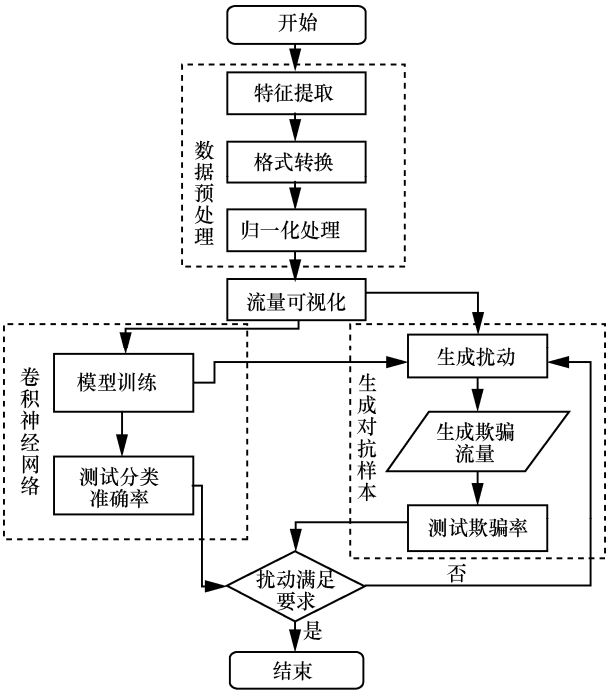


图 3 算法 1 流程

接下来，对攻击者选用的卷积神经网络进行构建和训练 Training(TF, mode)，使其能对可视化后的流量数据进行分类并测试分类的准确率；同时，选择扰动生成方法，将训练好的卷积神经网络模型的结构和参数作为已知条件，产生各自的扰动 CraftingPerturbation (method)，与真实流量叠加后形成欺骗流量 TD=Generate AdvSample()，并以可视化的方法比较欺骗流量和真实流量的区别；最后对生成的欺骗流量使用训练的卷积神经网络进行分类测试 Evaluate(TD)，并将测试结果与对真实流量分类的准确率进行比较。

本文设计的欺骗流量对抗样本生成算法的时间复杂度主要由三部分组成，分别为流量数据集的预处理（包括特征值提取、格式转换和归一化处理）、卷积神经网络的训练和测试，以及欺骗对抗流量的生成。其中，流量数据集的预处理部分的时间复杂度与流量数据集的特征的种类数量有关，本文将特征值视为图像的像素值，特征的种类数量为流量可视化后图像横纵像素数量的乘积，如本文中使用的流量数据集具有 248 个特征，将流量数据可视化后形成 16×16 的图像（248 个特征外的用 0 补足），假设流量特征的数量为 f，形成的可视化流量图像的大小为 x^2，因为对图像像素处理的时间复杂度为 $O(x^2)$，则预处理的时间复杂度为 $O(x^2 \approx (\sqrt{f})^2 = f) = O(f)$，即该部分的时间复杂度与流量特征的数量呈线性关系。卷积神经网络的训练和测试部分的时间复杂度为 $O\left(\sum_{l=1}^{d} n_{l-1} s_l^2 n_l m_l^2\right)$[30]。其中，$l$ 是神经网络的第 l 个卷积层；d 是神经网络的深度；m 是每个卷积核输出特征图的边长；s 是每个卷积核的边长；n 是每个卷积核的通道数，即输入通道数。对于欺骗对抗流量生成部分，由于其与卷积神经网络的训练和测试部分同时进行，时间消耗远小于卷积神经网络训练和测试所需要的时间。因此，在流量数据集特征数量一定的情况下，$O(f)$ 可忽略不计，算法 1 的时间复杂度为 $O\left(\sum_{l=1}^{d} n_{l-1} s_l^2 n_l m_l^2\right)$。

4 实验分析

本文构建如图 4 所示的攻防场景。假设攻击者

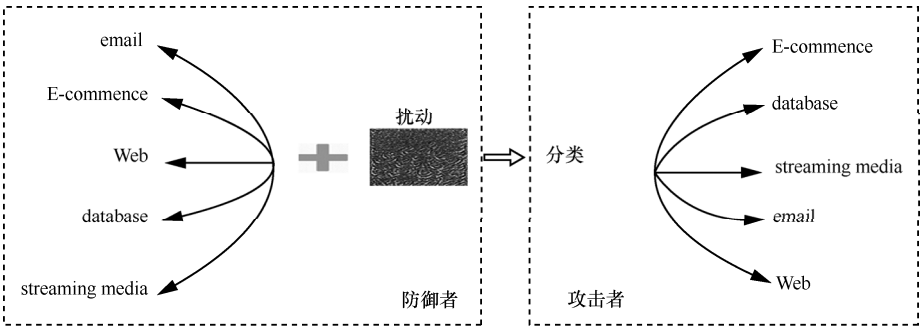


图 4 攻防场景

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

能够观测到 2 个主机节点之间的不同应用的 flow 级流量，使用基于深度学习的分类模型进行流量分类以发起进一步攻击；防御者使用本文提到的欺骗流量对抗样本生成方法，在正常流量中添加不同的扰动，从而使攻击者在进行流量分类攻击时出现分类错误，达到欺骗的目的。

实验环境和参数如表 4 所示。

表 4　　实验环境和参数

环境	参数
操作系统	Windows 10,64 bit
处理器	Intel Core i78706G, 3.1 GHz
内存	8 GB, LPDDR3, 2 133 MHz
Pycharm	Community Edition 11.05
Tensorflow	3.7 版本
训练集数量/个	65 036
测试集数量/个	23 801

4.1　数据集

本文使用的数据集为流量分类常用的 Moore 数据集[31]，作为图 4 中攻击者观测到的 flow 级流量。该数据集收集了 2003 年 8 月 20 日某大学 1 000 多名研究人员和工作人员使用的千兆以太网链路上的流量数据，通过抽样算法从每一条完整的 TCP 双向流抽样得到 377 526 条网络样本，包含 248 项属性，这些属性作为 3.1 节中攻击者模型的特征集合 X。经人工分类，将每个 flow 分类为 12 类应用类别中的某一类，其格式如图 5 所示，应用类别信息如表 5 所示。本文实验中使用 65 036 条记录作为训练集，23 081 条记录作为测试集，对应攻击者模型中的正常流量 TF 。

4.2　预处理

Moore 数据集是 arff 文件格式，需要对其按算法 1 进行预处理和格式转换 TransArfftoIDX。首先，将 arff 文件中的每一条流量记录按图 5 格式读出，前 248 项特征值作为正常流量，最后一项所属应用分类作为其对应的标签值。其次，为消除流量之间的量纲关系，使数据之间具有可比性，对数据集所有流量的每一项特征进行数值归一化。最后，归一化后的流量记录有 248 项特征，将其构建为 16×16 的矩阵，矩阵的后 8 行和后 8 列填充 0。将构建后的矩阵元素作为像素点，元素值即为像素灰度值，可以用 16×16 的灰度图像可视化流量。图 6(a)为不同应用的流量可视化；图 6(b)分为 4 组，即 Web、attack、FTP-data、email，每一组为同一应用但不同记录的流量的可视化，不同种类流量之间的可视化效果区分度比较明显。

表 5　　Moore 数据集应用类别信息

应用类别	flow 记录数量/条	所占比例	对应标签
Web	328 092	86.906%	0
email	28 567	7.567%	1
FTP-control	3 054	0.809%	2
FTP-pasv	2 688	0.712%	3
attack	1 793	0.475%	4
P2P	2 094	0.555%	5
database	2 648	0.701%	6
FTP-data	5 797	1.536%	7
multimedia	5 76	0.153%	8
services	2 099	0.556%	9
interactive	110	0.028%	10
games	8	0.002%	11
合计	377 526	100%	12

4.3　攻击者分类模型

本文使用的攻击者分类模型为 LeNet-5 卷积神经网络模型，其包括输入层、卷积层、池化层、卷积层、池化层、全连接层和输出层，广泛应用于对网络流量应用的分类[32]。本文根据文献[33]改进 LeNet-5 卷积神经网络模型，如图 7 所示，参数如表 6 所示。输入层设计为 16×16 矩阵，与可视化后的流量相对应；输出层为 12 个神经元，与分类结果相对应，即 3.1 节攻击者模型的应用类型集合 C 。

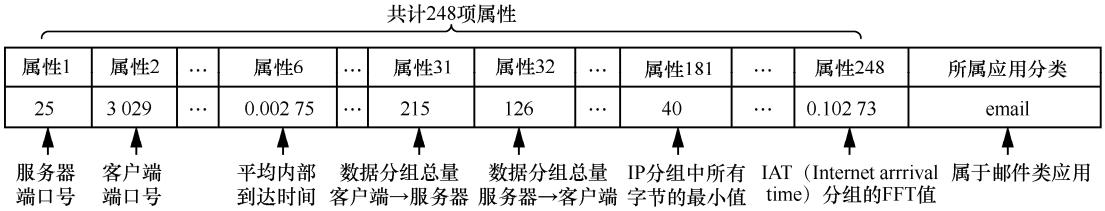
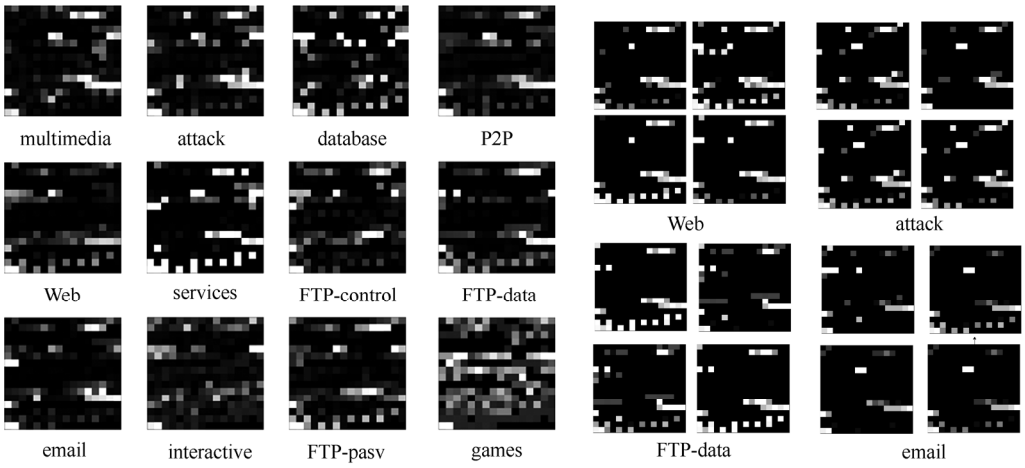


图 5　Moore 数据集流量记录的格式

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net



(a) 不同应用的流量可视化　　(b) 同一应用不同记录的流量可视化

图 6　不同应用类型的流量可视化和相同应用不同记录的流量可视化

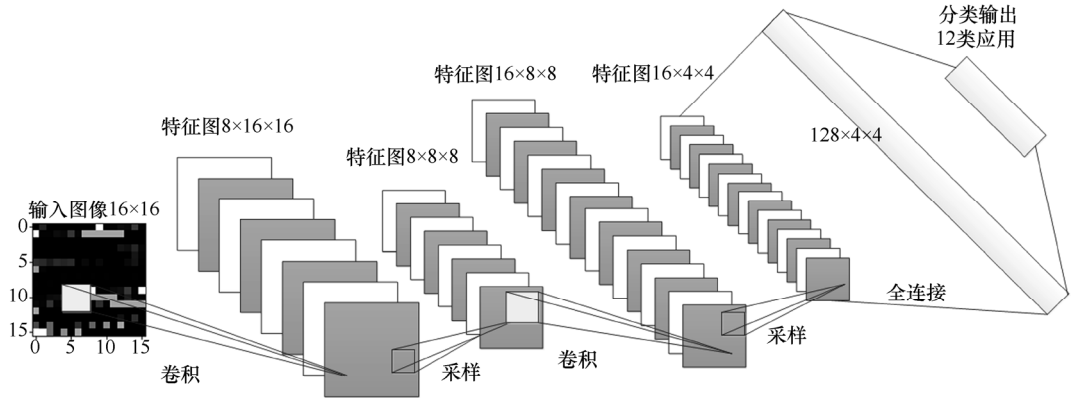


图 7　改进的 LeNet-5 卷积神经网络模型

表 6　　LeNet-5 网络的网络结构及参数

	名称	参数
	输入层	16×16
C1 卷积层	卷积核	8×(3×3)
	输出	8×(16×16)
S2 池化层	采样窗口	2×2
	输出	8×(8×8)
C3 卷积层	卷积核	16×(5×5)
	输出	16×(8×8)
S4 池化层	采样窗口	2×2
	输出	16×(4×4)
C5 全连接层	卷积核	128×(4×4)
	输出	128×1
	输出层	12×1

4.4　生成扰动

本节根据算法 1 和 3 种生成扰动的方法，在训练好的卷积神经网络模型基础上按照生成的 3 种扰动形成防御者模型的欺骗流量 TD，以 Web 应用为例，不同方法生成的扰动如图 8 所示。图 8(a)为 Web 应用原始流量不同记录的可视化图像，图 8(b)～图 8(c)为不同扰动生成方法产生扰动后叠加到原始流量形成的欺骗流量的可视化图像。从图 8 可以非常直观地发现，增加扰动后的欺骗流量与原始真实流量的可视化区别，且采用 FGSM 生成的样本需要的扰动最明显。

4.5　实验结果及分析

实验结果的评价指标包括以下两部分：1) 分类欺骗率，包括总体欺骗率（GFR, general fooling rate）和单类欺骗率（SCFR, single classification fooling rate）；2)欺骗流量对抗样本生成时间（DFSCT, deception flow sample crafting time）。相关计算式如下。

$$\begin{aligned} \mathrm{SCFR}_i &= 1-\mathrm{Acc}_i \\ \mathrm{Acc}_i &= \frac{\mathrm{TP}_i}{\mathrm{TP}_i+\mathrm{FN}_i} \end{aligned} \tag{4}$$

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

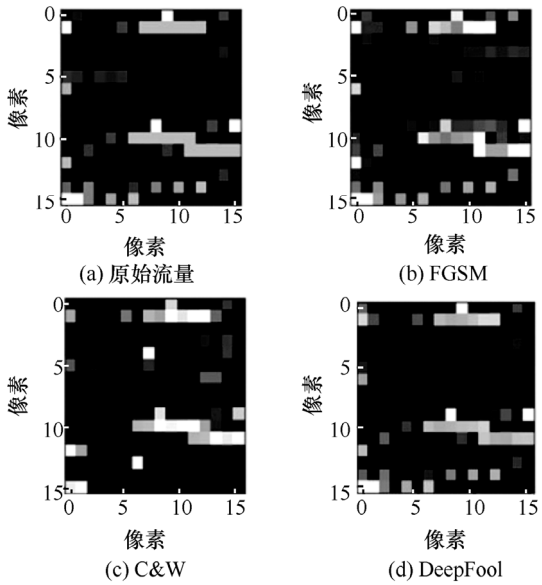


图 8　不同方法生成的扰动

$$\mathrm{GFR}=1-\mathrm{GAcc}$$

$$\mathrm{GAcc}=\frac{\sum_{i=1}^{m}\mathrm{TP_i}}{\sum_{i=1}^{m}(\mathrm{TP}_i+\mathrm{FN}_i)} \tag{5}$$

其中，Acc_i 是第 i 类样本的分类准确率，GAcc 是测试样本的整体分类准确率，TP_i 是实际类型为 i 的样本中被分类模型正常预测的样本数，FN_i 是实际类型为 i 的样本被分类模型误判为其他模型的样本数，$m=12$ 表示共有 12 类。总体欺骗率、欺骗流量对抗样本生成时间如表 7 所示。

表 7　不同方法的总体欺骗率和欺骗流量对抗样本生成时间

扰动生成方法	总体欺骗率	生成时间/s
FGSM	99.05%	143.61
DeepFool	97.38%	86.63
C&W	67.97%	60.84

未生成欺骗流量之前，攻击者利用 LeNet-5 对真实网络流量进行分类的准确率为 99.04%。针对 3 种不同的扰动生成方法，其对应的模型训练准确率、测试准确率和总体欺骗率如图 9 所示，可以看出，实施算法 1 之后，不论使用哪种扰动生成方法，流量的应用类型被错误分类的概率大大提高，以 FGSM 方法为例，攻击者使用 LeNet-5 对生成的欺骗网络流量进行分类时错误率达到了 99%以上。

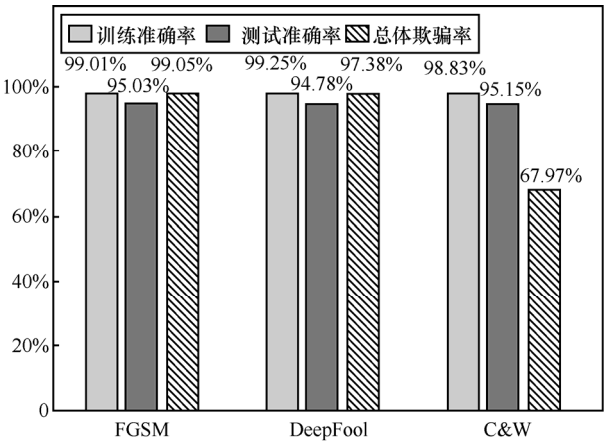


图 9　不同方法的训练准确率、测试准确率和总体欺骗率对比

根据式(5)，统计 3 种扰动生成方法对应的单类欺骗率和欺骗后的分类分布，如表 8 所示。

表 8 中，本文选择了样本数量前 5 位的流量应用，分类分布中描述了原始流量被分类成不同流量的分布情况，以 Web 流量为例，在使用 FGSM 时，22%被误分为 email 类，74%被误分为其他类，其他类为数据集中 Web、email、FTP-control、FTP-pasv、attack 之外的 7 类流量，对 Web 而言，其主要被误分为了其他中的 P2P 类。

以 C&W 方法为例，图 10 表示了各类流量应用生成的欺骗流量样本及被欺骗后的分类情况，图 10(a)表示不同的原始流量，图 10(b)表示 C&W 方法产生的扰动，图 10(c)表示形成的欺骗流量对抗样本，图 10 中的数字表示被欺骗后的错误分类的标签与比例。

从上述实验结果可以得出以下结论。1) 针对 LeNet-5 卷积神经网络模型，算法 1 选用的不同的扰动生成方法均起到了很好的欺骗效果，其中 FGSM 的欺骗效果最好但生成对抗样本的时间最长，DeepFool 方法的欺骗效果和生成时间居中，而 C&W 方法的欺骗效果不好但生成样本的时间最短，原因是流量分类的欺骗过程不像对抗图像分类过程那样对添加的扰动有严格的约束，不必形成“肉眼不可见”的扰动。2) 从对数据集的总体欺骗率来看，FGSM 和 DeepFool 方法显示出很高的对抗性，主要因为这 2 种方法是针对非特定目标的对抗样本生成，而 C&W 方法虽然在图像分类中具备针对特定目标进行欺骗的能力，但对非特定目标的欺骗能力相对其他 2 种方法略差。3) 对比不同扰动方法生成的欺骗流量对抗样本可知，采用 FGSM 生成的样本需要的扰动明显多于采用 DeepFool 和 C&W

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

表 8 不同方法的单类欺骗率与分类分布情况

流量应用	单类欺骗率		分类分布					
	方法	欺骗率	Web	email	FTP-control	FTP-pasv	attack	其他
Web	FGSM	96%	4%	22%	0	0	0	74%
	DeepFool	98%	2%	32%	1%	0	0	64%
	C&W	73%	27%	16%	0	0	0	56%
email	FGSM	96%	1%	4%	0	0	0	95%
	DeepFool	96%	14%	4%	0	0	0	82%
	C&W	68%	0	32%	0	0	0	68%
FTP-control	FGSM	98%	0	0	2%	46%	0	51%
	DeepFool	94%	0	0	6%	91%	0	3%
	C&W	61%	0	0	39%	61%	0	0
FTP-pasv	FGSM	94%	0	0	73%	6%	0	21%
	DeepFool	90%	0	0	88%	10%	0	2%
	C&W	74%	0	0	73%	26%	0	1%
attack	FGSM	98%	21%	12%	0	0	2%	65%
	DeepFool	98%	1%	0	0	0	2%	97%
	C&W	68%	44%	0	0	0	32%	24%

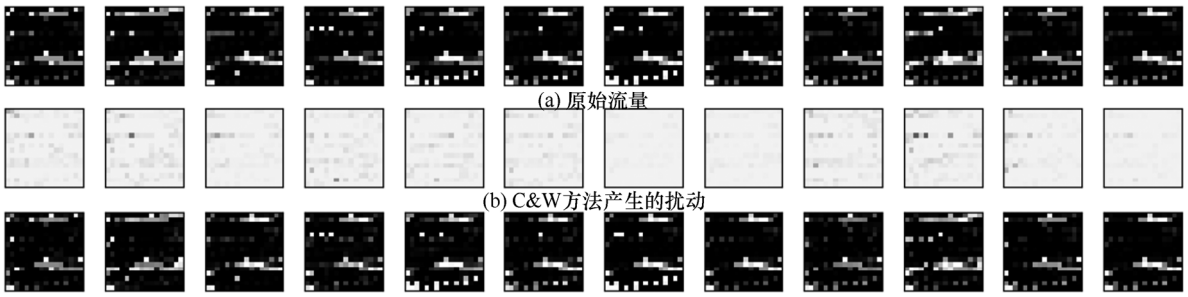


图 10 C&W 方法对各类流量应用生成的欺骗流量样本与分类情况

方法生成的样本所需扰动，验证了 DeepFool 和 C&W 方法所生成扰动的细微性。4) 从对流量的单类欺骗率来看，虽然 3 种方法的欺骗效果不同，但 3 种方法在将流量类型错误归类的趋势上大体相同，比如 Web 类和 P2P 类之间被错分的概率较大，services 类被误分为 database 的概率较大，3 种 FTP 类之间被错分的概率较大，这些都反映了流量种类之间的相似程度。

5 结束语

本文首先介绍了流量分类领域的研究现状，从防御者的角度，针对攻击者可能实施的流量分类攻击，在基于相关研究的基础上引入对抗样本的概念，提出了一种基于对抗样本的欺骗流量生成方法。然后，分别描述了攻击者和防御者模型，以 3 种常见的扰动生成方法进行实验，将 LeNet-5 卷积神经网络作为攻击者的分类模型，采用流量可视化的方法将正常网络流量转换为灰度图像，分别形成 3 种不同的欺骗流量对抗样本，对目标模型实施了欺骗。实验验证了所提方法的可行性。在将来的工作中，本文将针对以下 3 个方面进行更进一步的研究：1) 利用对抗样本的其他方法生成具有更高欺骗率的欺骗流量对抗样本，针对更复杂的深度神经网络进行欺骗；2) 充分考虑生成欺骗流量

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

对抗样本时的约束条件，研究针对特定目标的欺骗流量生成；3) 验证欺骗流量对抗样本对不同分类模型的迁移性，发现不同深度神经网络抵抗欺骗流量的稳健性。

参考文献：

[1] FRANK J. Artificial intelligence and intrusion detection: current and future directions[J]. Computers & Security, 1995, 14(1):31.

[2] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv Preprint, arXiv:1312.6199, 2013.

[3] SITAWARIN C, BHAGOJI AN, MOSENIA A, et al. Rogue signs: deceiving traffic sign recognition with malicious ads and logos[J]. arXiv Preprint, arXiv:1801.02780, 2018.

[4] YANG K, LIU J, ZHANG C, et al. Adversarial examples against the deep learning based network intrusion detection system[C]//2018 IEEE Military Communications Conference. Piscataway: IEEE Press, 2018: 559-564.

[5] HU W W, TAN Y. Generating adversarial malware examples for black-box attacks based on GAN[J]. arXiv Preprint, arXiv:1702.05983, 2017.

[6] 熊刚, 孟姣, 曹自刚, 等. 网络流量分类研究进展与展望[J]. 集成技术, 2012, 1(1): 32-42.
XIONG G, MENG J, CAO Z G, et al. Research progress and prospects of network traffic classification[J]. Journal of Integration Technology, 2012, 1(1):32-42.

[7] WANG Z. The applications of deep learning on tracidentification [J]. BlackHat USA, 2015,24(11): 21-26.

[8] WANG W, ZHU M, WANG J, et al. End-to-end encrypted traffic classification with one-dimensional convolution neural networks[C]//2017 IEEE International Conference on Intelligence and Security Informatics. Piscataway: IEEE Press, 2017: 43-48.

[9] RIMMER V, PREUVENEERS D, JUAREZ M, et al. Automated website fingerprinting through deep learning[J]. arXiv Preprint, arXiv:1708.06376, 2017.

[10] WANG W, ZHU M, ZENG X, et al. Malware traffic classification using convolutional neural network for representation learning[C]//2017 International Conference on Information Networking. Piscataway: IEEE Press, 2017: 712-717.

[11] PANCHENKO A, NIESSEN L, ZINNEN A, et al. Website fingerprinting in onion routing based anonymization networks[C]//Proceedings of 27 the 10th annual ACM workshop on Privacy in the electronic society. New York: ACM Press, 2011: 103-114.

[12] DYER K P, COULL S E, RISTENPART T, et al. Peek-a-boo, i still see you: why efficient traffic analysis countermeasures fail[C]//2012 IEEE Symposium on Security and Privacy. Piscataway: IEEE Press, 2012: 332-346.

[13] CUI W, YU J, GONG Y, et al. Realistic cover traffic to mitigate website fingerprinting attacks[C]//2018 IEEE 38th International Conference on Distributed Computing Systems. Piscataway: IEEE Press, 2018: 1579-1584.

[14] WANG T, GOLDBERG I. Walkie-talkie: an efficient defense against passive website fingerprinting attacks[C]//Proceedings of the 26th USENIX Security Symposium. Berkeley: USENIX Association, 2017: 1375-1390.

[15] DINGLEDINE R. Obfsproxy: the next step in the censorship arms race[R]. TOR Project official, (2012-05-23)[2020-03-20].

[16] MOGHADDAM H, LI B, DERAKHSHANI M, et al. SkypeMorph: protocol obfuscation for TOR bridges[C]//Proceedings of the 2012 ACM Conference on Computer and Communications Security. New York: ACM Press, 2012: 97-108.

[17] LI F F, KAKHKI A M, CHOFFNES D, et al. Classifiers unclassified: an efficient approach to revealing IP traffic classification rules[C]//Proceedings of the 2016 Internet Measurement Conference. New York: ACM Press, 2016: 239-245.

[18] 张思思, 左信, 刘建伟. 深度学习中的对抗样本问题[J]. 计算机学报, 2019, 42(8): 1886-1904.
ZHANG S S, ZUO X, LIU J W. The problem of the adversarial examples in deep learning[J]. Chinese Journal of Computers, 2019, 42(8): 1886-1904.

[19] GOODFELLOW I, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv Preprint, arXiv: 1412.6572, 2014.

[20] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial examples in the physical world[J]. arXiv Preprint, arXiv: 1607.02533 2016.

[21] MOOSAVI-DEZFOOLI S M, FAWZI A, FAWZI O, et al. Universal Adversarial Perturbations[C]// The IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 1765-1773.

[22] ATHALYE A, ENGSTROM L, ILYAS A, et al. Synthesizing robust adversarial examples[J]. arXiv Preprint, arXiv: 1707.07397, 2017.

[23] MOOSAVI-DEZFOOLI S M, FAWZI A, FROSSARD P. DeepFool: a simple and accurate method to fool deep neural networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016:2574-2582.

[24] METZEN J H, KUMAR M C, BROX T, et al. Universal adversarial perturbations against semantic image segmentation[C]//The IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 2755-2764.

[25] MOPURI K R, GARG U, BAHU R V. Fast feature fool: a data independent approach to universal adversarial perturbations[J]. arXiv Preprint, arXiv: 1707.05572, 2017.

[26] MOPURI K R, GANESHAN A, BABU R V. Generalizable data-free objective for crafting universal adversarial perturbations[C]//IEEE Transactions on Pattern Analysis and Machine Intelligence. Piscataway: IEEE Press, 2019,41(10): 2452-2465.

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

[27] VERMA G, CIFTCIOGLU E, SHEATSLEY R, et al. Network traffic obfuscation: an adversarial machine learning approach[C]//2018 IEEE Military Communications Conference. Piscataway: IEEE Press, 2018: 1-6.

[28] 潘文雯, 王新宇, 宋明黎, 等. 对抗样本生成技术综述[J].软件学报, 2020, 31(1): 67-81.
PAN W W, WANG X Y, SONG M L, et al. Survey on generating adversarial examples[J]. Journal of Software, 2020, 31(1): 67-81.

[29] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//2017 IEEE Symposium on Security and Privacy. Piscataway: IEEE Press, 2017: 39-57.

[30] HE K M, SUN J. Convolutional neural networks at constrained time cost[C]//The IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 5353-5360.

[31] MOORE A W, ZUEV D. Discriminators for use in flow-based classification[R]. Intel Research, Cambridge, (2005-08)[2020-03-19].

[32] LENCUN Y, BOTTOU L, BENGIO Y. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 862: 2278-2324

[33] 王勇, 周慧怡, 俸皓, 等. 基于深度卷积神经网络的网络流量分类方法[J]. 通信学报, 2018, 39(1): 14-23.
WANG Y, ZHOU H Y, FENG H, et al. Network traffic classification method basing on CNN[J]. Journal on Communications, 2018, 39(1): 14-23.

[作者简介]

胡永进（1981− ），男，山东潍坊人，信息工程大学讲师，主要研究方向为主动防御、态势感知。

郭渊博（1975− ），男，陕西周至人，博士，信息工程大学教授、博士生导师，主要研究方向为大数据安全、态势感知。

马骏（1981− ），男，山西阳泉人，博士，信息工程大学副教授，主要研究方向为态势感知与威胁发现。

张晗（1985− ），女，河南项城人，信息工程大学博士生，主要研究方向为自然语言处理、信息安全。

毛秀青（1980− ），男，安徽滁州人，信息工程大学副教授，主要研究方向为人工智能安全。

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net