

对抗性网络流量的生成与应用综述

王珏, 芦斌, 祝跃飞

引用本文

王珏, 芦斌, 祝跃飞. [对抗性网络流量的生成与应用综述](#)[J]. 计算机科学, 2022, 49(11A): 211000039-11.
WANG Jue, LU Bin, ZHU Yue-fei. [Generation and Application of Adversarial Network Traffic:A Survey](#)
[J]. Computer Science, 2022, 49(11A): 211000039-11.

相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

[基于机器学习的剩余使用寿命预测实证研究](#)

Empirical Research on Remaining Useful Life Prediction Based on Machine Learning
计算机科学, 2022, 49(11A): 211100285-9. <https://doi.org/10.11896/jsjcx.211100285>

[开放式环境下基于向量表征与计算的动态访问控制](#)

Vector Representation and Computation Based Dynamic Access Control in Open Environment
计算机科学, 2022, 49(11A): 210900217-7. <https://doi.org/10.11896/jsjcx.210900217>

[基于差分进化算法的字符对抗验证码生成方法](#)

Adversarial Character CAPTCHA Generation Method Based on Differential Evolution Algorithm
计算机科学, 2022, 49(11A): 211100074-5. <https://doi.org/10.11896/jsjcx.211100074>

[基于流量分析发现未知UDP反射放大协议](#)

Discovery of Unknown UDP Reflection Amplification Protocol Based on Traffic Analysis
计算机科学, 2022, 49(11A): 211000089-5. <https://doi.org/10.11896/jsjcx.211000089>

[深度神经网络的对抗攻击及防御方法综述](#)

Survey of Adversarial Attacks and Defense Methods for Deep Neural Networks
计算机科学, 2022, 49(11A): 210900163-11. <https://doi.org/10.11896/jsjcx.210900163>

对抗性网络流量的生成与应用综述

王 珏 芦 斌 祝跃飞

信息工程大学网络空间安全学院 郑州 450001
数学工程与先进计算国家重点实验室 郑州 450001
(1152808097@qq.com)

摘 要 人工智能技术的井喷式发展正在深刻影响着网络空间安全的战略格局,在入侵检测领域显示出了巨大潜力。最近的研究发现,机器学习模型有着严重的脆弱性,针对该脆弱性衍生的对抗样本通过在原始样本上添加一些轻微扰动就可以大幅度降低模型检测的正确率。学术界已经在图像分类领域对抗性图片的生成与应用进行了广泛而深入的研究。但是,在入侵检测领域,对于对抗性网络流量的探索仍在不断发展。在介绍对抗性网络流量的基本概念、威胁模型与评价指标的基础上,对近年来有关对抗性网络流量的研究工作进行了总结,按照其生成方式与原理的不同将生成方法分为 5 类:基于梯度的生成方法、基于优化的生成方法、基于 GAN 的生成方法、基于决策的生成方法以及基于迁移的生成方法。通过对相关问题的讨论,就该技术的发展趋势进行了展望。

关键词: 网络安全;入侵检测;机器学习;对抗样本

中图法分类号 TP393

Generation and Application of Adversarial Network Traffic: A Survey

WANG Jue, LU Bin and ZHU Yue-fei

School of Cyberspace Security, Information Engineering University, Zhengzhou 450001, China
State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou 450001, China

Abstract The spurt of artificial intelligence technology is profoundly affecting the strategic landscape of cyberspace security, showing great potential in the field of intrusion detection. Recent research finds that machine learning models have severe vulnerabilities, and the adversarial samples derived from this vulnerability can significantly reduce the correctness of model detection by adding some minor perturbations to the original samples. The generation and application of adversarial images has been extensively and intensively studied by academics in the field of computer vision. However, in the field of intrusion detection, the exploration of adversarial network traffic continues to evolve. Based on an introduction to the basic concepts, threat models and evaluation metrics of adversarial network traffic, the research works on adversarial network traffic in recent years are summarized, and the generation methods are classified into five categories according to their generation methods and principles: 1) gradient-based generation method; 2) optimization-based generation method; 3) GAN-based generation method; 4) decision-based generation method; 5) migration-based generation method. Through the discussion of related issues, an outlook on the development trend of this technology is presented.

Keywords Cybersecurity, Intrusion detection, Machine learning, Adversarial sample

1 引言

网络是现代社会基础设施的重要组成部分并且不断受到恶意攻击的威胁,入侵检测系统(Intrusion Detection System, IDS)在检测网络中的恶意行为方面起着至关重要的作用^[1]。近年来,得益于机器学习(Machine Learning, ML)理论的丰富以及相关方法的蓬勃发展,人工智能技术已在各行各业大放异彩,在入侵检测领域也显示出了巨大潜力。基于机器学习的 IDS 可以将网络入侵检测问题建模成一个针对网络流量的分类问题,然后由传统的机器学习或者深度学习(Deep Learning, DL)模型进行分类预测。随着网络流量的指数级增长和

许多数据密集型通信应用的出现,基于机器学习的 IDS 因为可以从大数据量的网络中提取特征的高维向量以检测入侵而受到了研究人员的广泛关注^[2-5]。

但是,最近的研究表明,机器学习模型可能容易受到对抗性攻击,对其检测能力产生负面影响。Szegedy 等^[6]发现,通过给图像添加肉眼难以分辨的扰动,来使已训练神经网络的分类错误的概率最大化,导致模型无法得到正确的分类结果,因此将这类添加了扰动的输入定义为对抗样本(Adversarial Example)。按照攻击者所掌握的模型相关信息,生成对抗样本的攻击策略主要分为白盒攻击与黑盒攻击两类。快速梯度符号法(Fast Gradient Sign Method, FGSM)^[7]、基本迭代法

基金项目:国家重点研发计划前沿科技创新专项基金(2019QY1300)
This work was supported by the Cutting-edge Science and Technology Innovation Project of the Key R & D Program of China(2019QY1300).
通信作者:祝跃飞(yfzhu17@sina.com)

(Basic Iterative Method, BIM)^[8]、基于雅可比的显著性图攻击(Jacobian-based Saliency Map Attack, JSMA)^[9]、DeepFool方法^[10]、C&W攻击^[11]等都是典型的白盒攻击策略,因为它们需要知道目标模型的内部信息。在攻击者只能获取模型的输入与输出的情况下,零阶优化(Zeroth Order Optimization, ZOO)^[12]、通用对抗性扰动(Universal Adversarial Perturbation, UAP)^[13]、生成对抗网络(Generative Adversarial Networks, GAN)^[14]、训练替代模型(Training Substitute Model)^[15]等黑盒攻击策略则更为有效。Fawazi等^[16]、Goodfellow等^[7]、Tabacof等^[17]都尝试对对抗样本产生的原因进行解释,但目前学术界还无法得到一个广泛的共识。

早期,针对对抗样本的生成与应用研究主要集中在图像分类领域。但最近,许多研究者开始将对抗样本技术应用到入侵检测领域,他们从流量维度出发尝试利用不同的方法生成对抗性网络流量(Adversarial Network Traffic, ANT),以逃避基于机器学习的流量分类与入侵检测,并取得了显著成效。

本文综述了对抗性网络流量的最新研究进展,并对该技术的发展趋势进行了展望。按照其生成方式和原理的不同,将对抗性网络流量的生成方法归纳为以下5类:1)基于梯度的生成方法;2)基于优化的生成方法;3)基于GAN的生成方法;4)基于决策的生成方法;5)基于迁移的生成方法。

本文第1节介绍了对抗性网络流量的基本概念、威胁模型与评价指标;第2节介绍了在图像分类领域已经取得成功的常用对抗性攻击策略;第3节介绍了对抗性网络流量的生成方法与应用;第4节对对抗性网络流量的相关问题进行了讨论,最后总结全文并展望未来。

2 对抗性网络流量

2.1 基本概念

2013年Szegedy等^[6]在图像分类领域发现了针对深度神经网络的攻击威胁并提出了对抗样本的概念。对抗样本(Adversarial Example)是指,在原数据集中通过人工添加

肉眼不可见或在经处理不影响整体的肉眼可见的细微扰动所形成的样本,这类样本会导致训练好的模型以高置信度给出与原始样本不同的分类输出^[18]。攻击者可能会得到训练模型(参数、结构)的知识,但是不被允许修改模型,这也是线上机器学习设备的一个共同假设。

与之类似,对抗性网络流量指通过对原始流量的内容或者统计特征添加扰动来实现检测逃逸的网络流量。与对抗性图像只需要满足在视觉上的不可分辨性不同,对抗性网络流量的生成条件更加苛刻。根据攻击原则和目的的不同,每类恶意流量都有其特定的功能性特征,为了保持与原始流量相同的可执行性和攻击性,对抗性网络流量在生成过程中必须将扰动添加的范围约束在由非功能性特征组成的可行域之内。除了僵尸网络与后门流量等少部分情况外,攻击者通常只能操控单方向的流量数据,并且将流量数据从输入空间到特征空间的映射过程是不可逆和不可微分的,因此对网络流量特征的提取/映射会更为复杂。2017年Rigaki^[19]首次尝试利用FGSM和JSMA对抗样本技术迁移到入侵检测领域。2018年Lin等^[20]提出了一种新的GAN框架IDSGAN,实现了针对IDS的黑盒对抗性攻击。2019年Liu等^[21]提出了一种基于遗传算法的对抗性网络流量黑盒生成方法,实现了匿名网络的抗网站指纹模型检测。2019年Ibitoye等^[22]利用BIM等白盒攻击策略,首次实现了针对物联网环境下IDS的对抗性网络流量检测逃逸。

2.2 威胁模型

假设攻击者从一系列具有恶意意图的网络流量开始,并希望通过不同的对抗性网络流量生成方法来逃避基于机器学习的流量分类与入侵检测,构建的威胁模型如图1所示。攻击者会根据不同的场景、假设和质量要求,在对抗性网络流量中决定他们需要的属性,然后部署特定类别的生成方法。不同类别的对抗性网络流量生成方法中既包含从图像分类领域进行迁移和拓展的攻击策略,也包含直接针对流量数据提出的新框架与新方法。

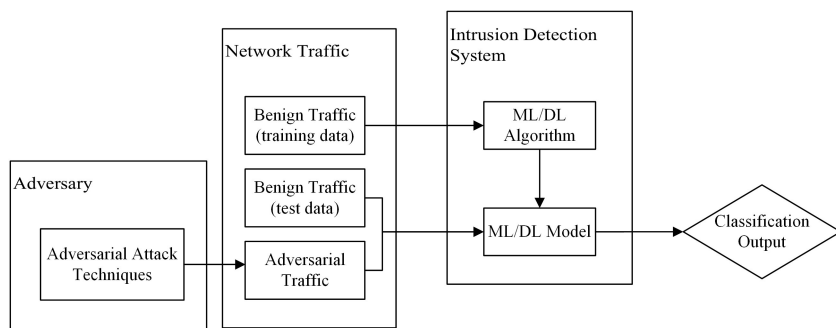


图1 对抗性网络流量威胁模型

Fig. 1 Adversarial network traffic threat model

我们采用Papernot等^[9]针对机器学习设备提出的威胁模型分类方法,从敌手知识(Adversarial Knowledge)和敌手目标(Adversarial Goals)两个维度对抗抗性网络流量的生成方法进行定性分类。

2.2.1 敌手知识

根据攻击者是否掌握关于模型的体系结构、训练数据、特征集合或者学习算法等相关信息,将对抗性网络流量生成方法分为白盒攻击与黑盒攻击。

(1)白盒攻击(White-box Attack):假设攻击者完全掌握

与目标模型相关的所有信息,包括参数、结构、激活函数等。在此假设下,攻击者可以通过求解模型输出关于输入的梯度来求解对抗性网络流量。

(2)黑盒攻击(Black-box Attack):假设攻击者对目标模型的了解非常有限,在此假设下,攻击者只能提供输入数据并查询模型的输出,但这通常是一种更贴近实际的情况,具有更现实的威胁性。

2.2.2 敌手目标

根据攻击者期望造成的安全破坏程度(完整性、可用性或

者隐私性)和攻击的专一性(针对性、非针对性),将对抗性网络流量的任务目标分为非目标攻击和目标攻击。

(1)非目标攻击(Non-targeted Attack):生成的对抗性网络流量让目标模型发生分类错误即可,并不指定最后模型的输出结果。非目标攻击很容易实施,因为它拥有更多的选择和空间来重定向输出。

(2)目标攻击(Targeted Attack):指定目标模型最后输出的结果,并尝试最大化目标攻击类的输出置信度。在某些特定的应用场景下,对抗性网络流量必须与背景流量的数据类别保持一致。因此,目标攻击具有更高的实施难度以及更大的威胁性。

2.3 评价指标

对抗性网络流量的评价指标主要包括扰动、迭代次数、成功率和鲁棒性。

2.3.1 扰动

对抗性网络流量能够成功实施欺骗的一个重要原因是与原始流量之间的差异足够小,因此添加的扰动不宜过大并且需要被限制在原始流量数据输入的可行域之内。对于扰动的评估主要采用范数的方式进行计算, l_p 表示 p 范数距离扰动的大小:

$$\|x\|_p = (\sum_{i=1}^n \|x_i\|_p)^{\frac{1}{p}} \tag{1}$$

其中, l_0 , l_2 和 l_∞ 是3种常见的计算方式。 l_0 用于计算原始流量中受到扰动的特征数量, l_2 用于计算对抗性网络流量与原始流量之间的欧氏距离, l_∞ 用于计算在所有受扰动特征中的最大扰动值。

2.3.2 迭代次数

在生成对抗性网络流量的过程中,白盒攻击与黑盒攻击都存在迭代次数。白盒攻击中的迭代次数主要指对特征或者数据包的修改次数,更少的迭代次数意味着更快速的对抗性网络流量生成速率,但往往也意味着更低的成功率;黑盒攻击中的迭代次数主要指对模型的访问次数,更少的迭代次数意味着更低的攻击成本,以及更小的被发现概率。因此,迭代次数通常是对抗性网络流量黑盒攻击的一个重要评价指标。

2.3.3 成功率

对抗性网络流量的生成主要是用于欺骗机器学习模型,因此可以从模型检测性能的降低程度来反映其攻击的成功率。对于模型检测的性能又通常可以利用准确率(Accuracy)、精确率(Precision)、召回率(Recall)、 F_1 值和ROC曲线等度量标准进行评估。在迭代次数有限的情况下,白盒攻击的成功率通常明显高于黑盒攻击的成功率。如何在迭代次数有限的条件下提高攻击成功率,是当前一个研究重点。

2.3.4 鲁棒性

对抗性网络流量的鲁棒性包含在真实环境中应用的鲁棒性和面对防御策略时的鲁棒性。一方面,由于生成的对抗性网络流量可能会受到噪声等因素的影响,如何抵御真实环境中的干扰是对抗性网络流量的一个重要指标;另一方面,随着近年来各种防御策略的提出,对抗性网络流量在面对这些防御策略时表现的鲁棒性也成为了一个重要指标。对抗性网络流量鲁棒性的高低通常以在不同应用场景下的攻击成功率进行衡量。

3 常用对抗性攻击策略

本节将根据攻击时需要的不同条件,分为白盒攻击与黑盒

攻击两类,对多种用于生成对抗样本的攻击策略进行介绍。这些攻击策略都在图像分类领域取得了巨大成功,为研究者将对抗样本技术迁移到入侵检测领域提供了重要的借鉴意义。

3.1 白盒攻击

3.1.1 快速梯度符号法

Goodfellow等^[7]认为,对抗样本是模型在高维流形空间过于线性化的结果,基于此结论提出了快速梯度符号法。FGSM仅沿着每个特征梯度符号的方向进行一步梯度更新,添加的扰动如下:

$$\eta = \epsilon \operatorname{sign}(\nabla \mathfrak{J}(\theta, x, y)) \tag{2}$$

其中, η 是需要添加的扰动, ϵ 是正则系数用于控制扰动 η 的大小, θ 是分类模型的参数, x 是模型的输入, y 是输入对应的正确标签, $\mathfrak{J}(\theta, x, y)$ 是训练神经网络时的损失函数。生成的对抗样本 x^{adv} 可以计算为 $x^{\text{adv}} = x + \eta$ 。

FGSM的优点是生成对抗样本只需要一次求导,计算速度非常快。但是,超参数 ϵ 取值的好坏严重影响攻击的成功率。

3.1.2 基本迭代法

Kurakin等^[8]采用类似梯度下降的方法,对FGSM进行拓展,进而提出了基本迭代法。BIM进行每一次迭代都仅生成很小的扰动,在经过多次的迭代后得到最终的结果,其迭代步骤如下:

$$x_0^{\text{adv}} = x$$

$$x_{n+1}^{\text{adv}} = \operatorname{Clip}_{x,\epsilon}\{x_n^{\text{adv}} + \epsilon \operatorname{sign}(\nabla \mathfrak{J}(\theta, x_n^{\text{adv}}, y))\} \tag{3}$$

其中, x_i^{adv} 表示经过第 i 次迭代后得到的对抗样本,每一次迭代更新的过程 $\operatorname{Clip}_{x,\epsilon}\{x'\}$ 可以定义为:

$$\operatorname{Clip}_{x,\epsilon}\{x'\} = \min\{255, x + \epsilon, \max\{0, x - \epsilon, x'\}\} \tag{4}$$

BIM通过增加损失函数值的大小迫使标签发生变化,但没有明确指定向哪个标签偏移,适合发动非目标攻击。

FGSM等同于迭代一次的BIM。之后陆续有许多基于FGSM或BIM的改进方法,Tramer等^[23]在FGSM的基础上增加了随机像素,提出了随机快速梯度符号法(rand-FGSM)。Madry等^[24]在BIM的基础上满足无穷范数的限制对初始点添加随机噪声,提出了投影梯度下降(Projected Gradient Descent, PGD)算法。

3.1.3 基于雅可比的显著性图攻击

Papernot等^[9]从图像分类领域的显著图中吸取灵感,提出了基于雅可比的显著性图攻击。JSMA基于显著性映射的概念生成对抗样本,使用对抗性显著图找到对目标模型特定输出影响程度最大的输入。首先计算关于模型输入 x 的Jacobian矩阵:

$$\mathfrak{J}_{F(x)} = \frac{\partial F(x)}{\partial x} = \left[\frac{\partial F_j(x)}{\partial x_i} \right]_{i \in 1, \dots, M, j \in 1, \dots, N} \tag{5}$$

其中, F 表示神经网络倒数第二层的输出(logits),其下标代表对应类别的logits; x_i 代表输入的第 i 个分量。

然后,文献^[9]定义了两种对抗性显著图的计算方式,以选择在每次迭代中要调整的特征。通过对每个样本平均修改4%的特征,获得了高达97%的攻击成功率。JSMA能通过添加少量的扰动来生成对抗样本,但是计算量太大,算法运行速度缓慢。

3.1.4 0DeepFool方法

Moosavi-Dezfooli等^[10]从几何分析出发,提出了DeepFool方法。定义 F 为分类问题的仿射平面, $F = \{x; f(x) =$

0}, 其中 $f(x) = \omega^T x + b$ 表示模型的分类函数。若要在平面上的某一点 x_0 处添加最小的扰动, 则扰动方向为点 x_0 垂直于平面 F 的方向, 大小为 $\Delta(x_0, f)$:

$$\begin{aligned} r_*(x_0) &:= \operatorname{argmin} \|r\|_2 \\ \text{s. t. } \operatorname{sign}(f(x_0 + \eta)) &= \operatorname{sign}(f(x_0)) = -\frac{f(x_0)}{\|\omega\|_2} \omega \end{aligned} \quad (6)$$

图 2 给出了 DeepFool 方法在线性二分类问题中的一个应用过程。

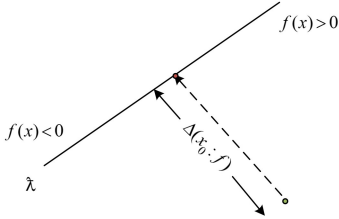


图 2 二分类问题中的 DeepFool 方法^[10]

Fig. 2 DeepFool method in binary classification problem^[10]

在面对非线性多分类问题时, 文献[10]没有直接求出每个分类面的最短距离, 而是每次计算点到非线性平面切面的距离, 然后只移动很小的一部分, 通过多次迭代慢慢逼近分类面。与 FGSM 和 JSMA 相比, DeepFool 方法改变的特征数量更少、添加的扰动更小, 但是计算量会随着分类的类别增多而增多, 不适合高维数据。

3.1.5 C&W 攻击

Carlini 等^[11]提出了 C&W 攻击, 这是一种在白盒攻击中攻击效率与攻击效果都非常好的方法。首先, 定义一种新的目标函数 g , 使得:

$$\begin{aligned} \min_{\eta} \|\eta\|_p + \epsilon \cdot g(x + \eta) \\ \text{s. t. } x + \eta \in [0, 1]^n \end{aligned} \quad (7)$$

当且仅当原始目标函数 $f(x^{\text{adv}}) = l'$ 时, $g(x^{\text{adv}}) \geq 0$, l' 表示目标对抗样本 x^{adv} 的分类标签。在这种情况下, 距离和惩罚项可以被更好地优化。作者提出的其中一种目标函数 g 可以定义为:

$$g(x') = \max(\max_{i \neq l'} (Z(x')_i) - Z(x')_{l'}, -k) \quad (8)$$

其中, Z 表示 softmax 函数, i 是原始样本的输出标签, k 是一个控制置信度的常数。然后, 又引入了新的变量 w 来代替原始变量 x , 实现对目标对抗样本的边界约束:

$$\begin{aligned} \min_{\eta} \|\eta\| + \epsilon \cdot g\left(\frac{1}{2} \tanh(w) + 1\right) \\ \text{s. t. } \eta = \frac{1}{2} \tanh(x) + 1 - x \end{aligned} \quad (9)$$

对于其中的正则系数 ϵ , 使用二分搜索方式来找到较为合适的值。

C&W 攻击是一种基于优化的方法, 可以突破现有的绝大部分对抗样本主动防御策略。但是, 与 FGSM 相似, 超参数 ϵ 和 k 设置的好坏对 C&W 攻击的成功率有重大影响。

3.2 黑盒攻击

3.2.1 零阶优化

Chen 等^[12]提出了一种基于零阶优化的攻击方式, 修改了 C&W 攻击中的 g 函数, 将其作为一个新的损失函数:

$$g(x') = \max(\max_{i \neq l'} (\lg[f(x)_i]) - \lg[f(x)]_{l'}, -k) \quad (10)$$

该损失函数对 C&W 攻击中的损失函数进行了对数处理, 使得该损失函数在目标函数中的占比更大, 并使用对称

差分商 (Symmetric Difference Quotient) 对梯度和 Hessian 矩阵进行估计:

$$\begin{aligned} \frac{\partial f(x)}{\partial x_i} &\approx \frac{f(x + he_i) - f(x - he_i)}{2h} \\ \frac{\partial^2 f(x)}{\partial x^2} &\approx \frac{f(x + he_i) + f(x - he_i) - 2f(x)}{h^2} \end{aligned} \quad (11)$$

其中, e_i 表示第 i 个分量为 1 且 h 为小常数的标准基向量。

因为这种攻击不要求已知梯度, 它可以直接部署到一个黑盒攻击中而不需要考虑模型迁移。但是, ZOO 的这种梯度估计是低效的, 它需要大量的计算成本来查询和估计梯度值。

3.2.2 通用对抗性扰动

Moosabi-Dezfooli 等^[13]在 DeepFool 方法的基础上加入了通用对抗性扰动的概念, 提出了一种计算通用扰动的方法, 构想出了一种通用的扰动向量并满足:

$$\begin{aligned} \|\eta\| &\leq \epsilon \\ P(x' \neq f(x)) &\geq 1 - \delta \end{aligned} \quad (12)$$

其中, ϵ 限制了通用扰动的大小; δ 控制了所有对抗样本的失败率。

对于每次迭代, 用 DeepFool 方法获得一个针对每一次输入的最小样本扰动, 并同时扰动更新为总扰动 v , 直到大多数数据样本都被欺骗 ($P < 1 - \delta$) 时, 停止循环。图 3 给出了一个通用扰动的迭代过程。

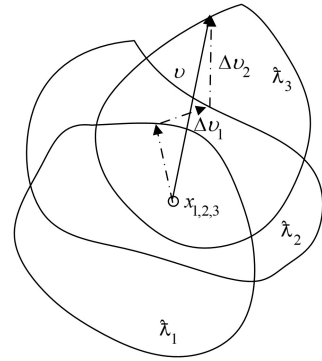


图 3 扰动迭代过程^[13]

Fig. 3 Perturbation iterative process^[13]

图 3 中, $x_{1,2,3}$ 是 3 个重合的样本点, $R_{1,2,3}$ 表示不同的分类域。先在 R_1 域中找到最佳的逃离分类域的扰动, 更新 x , 再根据新的 x 找到逃离 R_2 的扰动, 以此类推, 最终计算出扰动 v 。

UAP 可以对整个服从同一分布的数据集中的大部分样本进行干扰, 并对不同网络结构的模型均通用。

3.2.3 生成对抗网络

Goodfellow 等^[14]提出了一种通过对抗过程估计生成模型的新框架: 生成对抗网络。GAN 的核心思想来源于博弈论中的纳什均衡, 参与博弈的双方分别为一个生成器网络 (Generator Network) 和一个判别器网络 (Discriminator Network), 生成器通过学习真实的数据分布利用噪声生成对抗性数据, 而判别器尽可能去正确判别输入数据是来自真实数据还是来自生成器。在对抗训练的过程中, 双方都竭力优化自己的网络, 各自提高自己的生成能力和判别能力, 最终达到纳什均衡。

GNA 不需要针对应用的问题去设计专门的损失函数, 不需要显式地建模数据分布, 在不同领域都有着广泛的应用,

GAN 的结构示意图如图 4 所示。

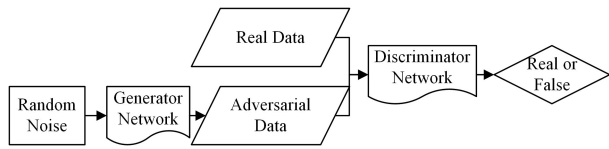


图 4 GAN 的结构示意图

Fig. 4 Structural diagram of GAN

3.2.4 训练替代模型

针对黑盒攻击假设, Papernot 等^[15]提出了训练替代模型 (Training Substitute Model) 的概念。具体来说, 攻击者可以给黑盒模型输入各种特征的样本, 通过观察不同样本所产生的模型输出的差异, 根据差异对模型的判定边界进行估计, 利用估计的边界生成一个替代模型, 由替代模型产生出原始黑盒模型的对抗样本, 该样本仍然能被原始模型分类错误。

文献^[15]针对黑盒假设下的深度神经网络 (Deep Neural Networks, DNN), 使用一个替代的神经网络进行拟合, 然后利用 FGSM 和 JSMA 生成对抗样本, 使得 DNN 错误分类了 84.24% 的对抗性输入。训练替代模型的方法在小数据集训练的模型上表现良好, 然而在面对边界更加复杂的数据集时, 其成功率会发生显著下降。

4 对抗性网络流量的生成与应用

在入侵检测领域, 攻击者已经成功利用多种对抗性网络流量生成方法实现了检测逃逸。我们按照其生成方式和原理的不同, 将对抗性网络流量的生成方法归纳为 5 类: 1) 基于梯度的生成方法; 2) 基于优化的生成方法; 3) 基于 GAN 的生成方法; 4) 基于决策的生成方法; 5) 基于迁移的生成方法。

4.1 基于梯度的生成方法

基于梯度的生成方法仅适用于白盒攻击。攻击者在完全掌握目标模型参数、结构等信息的前提下, 通过在梯度的反方向上添加扰动来增大新样本与原始样本的决策距离, 以生成对抗性网络流量。目前已经有研究者利用 FGSM, BIM, PGD 和 JSMA 等图像分类领域的攻击策略成功实现了针对不同场景下 IDS 的检测逃逸。

Rigaki^[19]首次尝试将对抗样本技术引入入侵检测领域, 使用 FGSM 和 JSMA 两种基于梯度的攻击策略, 利用多层感知机 (Multilayer Perceptron, MLP) 对 NSL-KDD 数据集^[25]进行训练以生成对抗性网络流量, 并选择了 4 种机器学习模型作为攻击目标: 决策树 (Decision Tree, DT)、随机森林 (Random Forest, RF)、线性支持向量机 (Linear Support Vector Machine, LSVM) 以及综合前 3 种模型的投票方法 (Voting Method)。实验重点评估了基于 JSMA 生成的对抗性网络流量对上述 4 种模型的影响, 从准确率、 F_1 值和 ROC 曲线 3 个维度进行评估发现: LSVM 受到的影响最大, 3 个指标平均降低 28%; RF 则表现出了最好的鲁棒性, 3 个指标平均只降低 3%, 这样的实验结果与在图像分类领域中得出的一些研究结论正好相反。最后, 他们还指出: FGSM 会使得原始流量数据的全部特征都发生微小的改变, JSMA 则平均只修改了 6% 的特征, 这意味着 JSMA 可能更适合运用到入侵检测领域, 充分说明了研究不同应用领域并满足其特定需求的重要性。

软件定义网络 (Software Defined Network, SDN) 是一种新型网络架构。其基本思想是将网络的数据层和控制层

分离, 通过软件编程在逻辑上实现对网络的集中控制与管理。Huang 等^[26]针对深度学习入侵检测设备进行了基于 SDN 的对抗性攻击实验, 收集了 SDN 控制器接收到的消息和统计报告中的数据, 以创建基于 SDN 的端口扫描攻击数据集。将 FGSM, JSMA, JSMA-RE 应用于 3 种不同的深度学习模型: 多层感知机、卷积神经网络 (Convolutional Neural Network, CNN)、长短期记忆网络 (Long Short-Term Memory, LSTM)。实验结果表明, 基于 JSMA 生成的对抗性网络流量对深度学习模型的影响较大, 模型检测的准确率降低了 14%~42%, 基于 FGSM 生成对抗性网络流量的用时最短但攻击效果也最不理想。这表明 JSMA 可能更适用于扰动测试数据, 而 FGSM 则更适用于扰动训练模型, 这与 Rigaki^[19]的研究结论具有相似性。

Ibitoye 等^[22]在物联网 (the Internet of Things, IoT) 环境下研究了对抗样本对基于深度学习的 IDS 的影响。他们选择在专用物联网环境中创建的 BoT-IoT 数据集^[27], 利用 FGSM, BIM 和 PGD 3 种对抗性攻击策略对自归一化神经网络 (Self-normalizing Neural Network, SNN) 和前馈神经网络 (Feed-forward Neural Network, FNN) 发起攻击。初始准确率高达 95.1% 的 FNN, 在遭受 FGSM 对抗样本攻击时准确率降为 24%, 利用 BIM 和 PGD 对抗样本重复实验后准确率分别降为 18% 和 31%。他们基于准确率、召回率、 F_1 值等多个度量标准比较了 SNN 和 FNN 的模型性能, 发现在多次实验中 FNN 的各项指标始终优于 SNN。但是, 相比 FNN, 在面对不同的对抗样本攻击时 SNN 具有更强的鲁棒性。最后, 还指出: 在物联网环境下, 数据集的特征规范化会对基于深度学习的 IDS 的鲁棒性产生负面影响; 当输入数据的特征规范化时, IDS 虽然具有更好的检测性能, 但是也更容易遭受对抗样本的威胁。

Papadopoulos 等^[28]也利用 BoT-IoT 数据集研究了物联网环境下基于机器学习的 IDS 的鲁棒性问题。在检测阶段, 他们利用 FGSM 针对人工神经网络 (Artificial Neural Networks, ANN) 生成对抗样本并发动了目标攻击与非目标攻击。研究发现, 随着正则系数 ϵ 逐渐增加到 1, 模型的准确率、精确率、召回率和 F_1 值等性能指标都发生了明显降低的现象; FGSM 非目标攻击比目标攻击具有更显著的负面影响, 在逃逸检测的同时会增加 IDS 的误报率。但是, FGSM, BIM 和 PGD 都会沿着每个特征梯度符号的方向进行梯度更新, 而攻击者不可能以如此细粒度的方式操纵流量数据, 也无法保证修改后的网络流量仍然保持相同的功能与可用性, Papadopoulos^[28]与 Ibitoye^[22]似乎都没有考虑到这方面的问题。

4.2 基于优化的生成方法

基于优化的生成方法即适用于白盒攻击也适用于黑盒攻击。攻击者将针对目标模型寻找的最小扰动转化为一个优化问题, 然后利用不同的策略寻找最优解以生成对抗样本。在入侵检测领域, 研究者除了将 DeepFool, C&W 攻击, ZOO 和 UAP 等攻击策略进行迁移, 还提出了一种基于遗传算法的对抗样本生成方法。

Zheng^[29]在 NSL-KDD 数据集上研究了基于深度学习的 IDS 的脆弱性问题, 选择了 4 种不同的对抗性攻击策略用于攻击多层感知机, 既有基于梯度的生成方法, 即 FGSM 和 JSMA, 也有基于优化的生成方法, 即 DeepFool 和 C&W 攻击。

从 ROC 曲线这个维度进行评估发现,对于原始数据集 MLP 预测的平均 AUC 值为 0.94,在 JSMA 攻击下所有数据类别的 AUC 值都低于 0.5;当发动 FGSM 非目标攻击时,Normal 类和 Dos 类的 AUC 值分别进一步降低至 0.21 和 0.15,而其余数据类别的平均 AUC 值保持在 0.5 左右;采用 DeepFool 方法时可以获得类似的效果,Normal 类和 Dos 类的 AUC 值分别为 0.15 和 0.14,但是其余数据类别的平均 AUC 值回升到 0.7 左右;FGSM 目标攻击获得的效果最明显,所有数据类别的平均 AUC 值仅为 0.44;而 C&W 攻击获得的效果最差,所有数据类别的平均 AUC 值高达 0.8。但是,Zheng^[29]最后也发表评论:由于攻击者所能操纵的资源有限,JSMA 在可用性和适用性方面会更具有吸引力;C&W 攻击的破坏性虽然不如其他 3 种方法,但当面对一些最先进的防御策略时会表现出更好的鲁棒性。

Hu 等^[30]也利用 FGSM,DeepFool 和 C&W 攻击生成对抗样本,对 LeNet-5 深度卷积神经网络实施了欺骗。他们选择了 Moore 数据集^[31]在 Flow-level 的 12 种应用类别共 88117 条记录来生成对抗性网络流量。实验结果表明,3 种不同的扰动生成方法均起到了很好的欺骗效果。在未生成对抗性网络流量之前,LeNet-5 对真实网络流量进行分类的准确率为 99.04%,FGSM 的总体欺骗率最高为 99.05%,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,C&W 攻击的总体欺骗率最低仅为 67.97%,但生成对抗样本的时间最短。他们还发现,虽然采取的是无目标攻击,但是 3 种方法在将不同应用类别错误归类的趋势上大体相同,反映了流量种类之间的相似程度。但是,将形成上述实验结果的原因归结为流量分类的欺骗过程,不需要像对抗图像分类过程那样对添加的扰动进行严格约束的观点并不完全正确,没有考虑到网络流量修改后的正常通信与攻击语义问题。

Hartl 等^[32]选择 CIC-IDS2017 数据集^[33]和 UNSW-NB15 数据集^[34]从 Packet-level 和 Flow-level 两个流量层面分别对循环神经网络(Recurrent Neural Networks,RNN)的可解释性和对抗鲁棒性进行了研究。首先,定义了一种新的评价指标,即对抗鲁棒性评分(Adversarial Robustness Score,ARS),用于评估 FGSM,PGD,C&W 攻击的效能,实验结果表明 C&W 攻击提供了最好的效能,FGSM 的效能最差。其次,引入了特征敏感度(Feature Sensitivity)的概念,用于量化分析某一种特征在多大程度上可以影响模型的预测结果,并拓展了现有的解释性方法尝试进一步深入理解 RNN 模型。实验结果表明,模型预测的结果主要受网络流中前几个数据包的影响,在后续的时间步骤中修改特征基本不会再改变其置信度。最后,还通过实验说明了部署对抗性训练程序(Adversarial Training Procedure)可以有效缓解来自对抗性网络流量的威胁。

Sadeghzadeh 等^[35]评估了一维卷积神经网络(One-Dimensional Convolutional Neural Networks,1D-CNN)在面对对抗性网络流量时的鲁棒性。他们选择 ISCXVPN2016 数据集^[36]基于 UAP 生成对抗性网络流量,从数据包分类(Packet Classification,PC)、数据流内容分类(Flow Content Classification,FCC)和数据流时间序列分类(Flow Time Series Classification,FTSC)3 种不同的输入空间发起黑盒攻击。针对 3 种

输入空间,提出了 3 种应用 UAP 的攻击策略,即 AdvPad 攻击、AdvPay 攻击和 AdvBurst 攻击,利用准确率、召回率和 F_1 值 3 个指标对模型的性能进行评估。实验结果表明,1D-CNN 在所有输入空间中鲁棒性都很低,模型性能会因为微小的扰动而显著降低;AdvPay 攻击对网络流量施加的带宽开销最小,并且显著降低了所有被测数据类别的召回率;在面对不同的防御策略时,AdvBurst 攻击则可能会表现出最好的鲁棒性。此外还发现,1D-CNN 对数据包有效载荷开始时的信息比结束时的信息更敏感,这与 Hartl^[32]的研究结论相似。

泊松隐马尔可夫模型(Poisson Hidden Markov Model)网站指纹攻击^[37]可以对捕获的数据包执行模式匹配,以检测数据流所属的网站,这使得匿名网络失去了它的作用。为了绕过该模型的检测,Liu 等^[21]提出了一种基于遗传算法的对抗样本黑盒生成方法。该方法针对的数据类型是流量数据,选择 Alexa-200 数据集^[37]进行实验。对于一个输入样本 X , $F(X)$ 表示输入样本 X 被正确识别的概率,该方法的目的是添加微小的扰动 η ,使得恶意样本 $X+\eta$ 的分类结果 $F(X+\eta) < 1 - F(X)$,并且扰动 η 越小越好。在这个过程中,需要将被扰动特征的数量最小化,每个特征的扰动程度也要最小化。为了确保扰动后流量的“有效性”,该方法对被扰动特征的类型和幅度进行了限制。同时,使用 l_2 范数来计算对抗样本与原始样本之间的欧氏距离。在确定扰动成功标准和扰动的度量方式后,结合遗传算法来选择特征的扰动程度以搜索最优解,从而实现了针对 PHMM 的黑盒对抗样本生成。假设原始的流量向量为:

$$X_i = [X_i^0, X_i^1, X_i^2, \dots, X_i^n] \quad (13)$$

对应的遗传算法中的个体向量为:

$$P_i = [P_i^0, P_i^1, P_i^2, \dots, P_i^n] \quad (14)$$

在遗传算法中适应度函数的计算式如下:

$$F(P_i) = \omega_d D(P_i) + \omega_e (1 - E(P_i)), j \in [0, n] \quad (15)$$

其中, ω_d 和 ω_e 是两个经验值,用于衡量两个变量的权重; $D(P_i)$ 衡量原始样本和对抗样本之间差异性的大小; $E(P_i)$ 表示生成的对抗样本在当前威胁模型下的攻击成功率。实验结果表明,该方法生成了高质量的对抗性网络流量,攻击成功率高达 97.16%。

4.3 基于 GAN 的生成方法

基于 GAN 的生成方法主要适用于黑盒攻击。生成器通过修改流量中的某些特定特征来生成对抗性网络流量,判别器则模拟黑盒模型辅助训练。目前,已经有许多基于 GAN 的衍生框架被提出并应用到入侵检测领域。

Pan 等^[38]基于 GAN 提出了一种新的框架,用于生成对抗样本,该框架分为生成模块、判别模块和验证模块。与传统的 GAN 仅以随机噪声为生成器输入不同,新框架的生成器输入直接是经过处理的恶意流量样本,用于解决生成器在优化过程中无法收敛的问题。还提出了弱相关位的概念,通过按位与操作将允许修改的数据范围约束在弱相关位的可行域之内,保证对抗性网络流量的可执行性和攻击性不被破坏,并选择 UNSW-NB15 数据集中的缓冲区溢出漏洞攻击流量进行实验。但是,初始准确率为 94% 的 AlexNet 检测器在面对对抗性网络流量时,检测准确率仍然有 84% 左右,欺骗效果不佳。

针对 GAN 在训练过程中存在的梯度消失问题,Arjovsky

等^[39]提出了使用权重裁剪(Weight Clipping)策略的 Wasserstein GAN(W-GAN)。基于 W-GAN, Lin 等^[20]提出了一种新的框架 IDSGAN,在流量维度上实现了针对 IDS 的黑盒对抗性攻击。由于 GAN 框架只适用于二分类问题,IDS-GAN 的生成器每次只能为 NSL-KDD 数据集中的某一种数据类别生成对抗性网络流量。为了全面深入地进行评估,针对支持向量机、朴素贝叶斯(Naive Bayes, NB)、多层感知机(Multi-layer Perceptron, MLP)、逻辑回归(Logistic Regression, LR)、决策树(Decision Tree, DT)、随机森林和 K 近邻(K-Nearest Neighbor, KNN)等多种模型发动了黑盒攻击。为了保持流量的功能特性,在博弈的过程中生成器仅通过修改流量中的部分非功能性特征来生成对抗性网络流量。实验结果表明,基于 IDSGAN 生成的对抗性网络流量可以使得所有模型的检测准确率都降低到 1% 以下,体现了极强的泛化能力。

Qiao 等^[40]也在 W-GAN 的基础上提出了一种可以将拒绝服务(Denial of Service, DoS)攻击流量伪装成正常流量的 DoS-WGAN 框架,该框架由生成器、转换器和判别器组成。DoS 攻击流量的伪装过程如下:1)对 KDDCup99 数据集^[41]中 DoS 类别的流量特征进行分类,将那些无法修改的特征和修改后会改变流量功能的特征标记为“unchanged”,其余特征标记为“changed”;2)生成器利用随机噪声生成扰动 $G(z)$;3)转换器将原始 DoS 攻击流量样本中标记为“unchanged”的特征与 $G(z)$ 中标记为“changed”的特征相结合,生成对抗性网络流量 $C(G(z), X_{DoS})$;4)判别器尝试区分正常数据和对抗样本以进行辅助训练;5)通过对 DoS-WGAN 的对抗训练,最终生成近似于正常流量的对抗性 DoS 攻击流量。将生成的对抗性网络流量输入到卷积神经网络(Convolutional Neural Network, CNN)中,模型检测的精确率从 97.3% 降低到 47.6%,验证了 DoS-GAN 在流量伪装方面的有效性。

Usama 等^[42]则选择 KDDCup99 数据集中 Probe 类的流量从攻防两个方面进一步研究了基于 GAN 的模型检测对抗。在攻击方面,借鉴了 Lin 等^[20]的实验思路,利用 GAN 框架对 Probe 类的非功能性内容特征进行修改,以生成对抗性网络流量;在防御方面,提出了一种基于 GAN 的防御训练机制,防御模型不仅根据输入数据进行训练,还根据生成器生成的对抗性数据进行训练,提高了 IDS 在面对对抗性攻击时的鲁棒性。将深度神经网络、逻辑回归、支持向量机、K 近邻、朴素贝叶斯、随机森林、决策树和梯度提升(Gradient Boosting, GB)作为黑盒 IDS 进行了攻击与对抗实验。

4.4 基于决策的生成方法

基于决策的生成方法即适用于白盒攻击也适用于黑盒攻击。攻击者向目标模型不断提供输入数据并查询输出,借助模型的决策反馈(判定流量正常/恶意),迭代地修改原始输入以生成对抗性网络流量。

Mohammad 等^[43]利用一种基于决策的对抗性网络流量生成方法选择 CIC-IDS2017 数据集对 3 种不同的 IDS 实施欺骗:在 Packet-level 欺骗了 Kitsune,在 Flow-level 欺骗了 DAGMM 和一种基于 BiGAN 的异常检测器。假设在本地部署有目标 IDS 的副本并知道其所有参数,采用了 3 种通用的数据包操作技术以生成对抗性网络流量:1)拆分数据包的有效载荷;2)调整数据包的发送速率;3)构造虚假数据包。基于本地 IDS 副本对输入数据的决策,不断地在数据包层面或者

数据流层面对网络流量进行裁剪,在不破坏底层网络协议与攻击语义的前提下最终实现了检测逃逸。通过上述裁剪过程,基于数据包特征训练的 IDS 的检测率降低了 70%,基于数据流特征训练的 IDS 的检测率降低了 68%,提高了入侵的成功率。

Aiken 等^[44]也采取了与 Mohammad 类似的方法,进行了基于 SDN 的入侵检测对抗性研究。选择 DARPA2009 数据集^[45]中的 SYN flood 类别的流量,对 SDN 环境的下逻辑回归、随机森林、支持向量机和 K 近邻等模型发起分布式拒绝服务(Distributed Denial of Service, DDoS)攻击。通过对 DDoS 攻击行为的分析,同样提出了 3 种对数据包的扰动操作:增加数据包载荷、降低数据包发送速率和构造双向流量。此外,还开发了一个对抗性评估工具 Hydra,基于黑盒假设向目标模型发起攻击并依据决策判定的结果对输入不断进行调整,针对不同目标模型确定最佳的扰动组合。实验结果表明,其提出的方法对所有被测模型都有不同程度的影响,具有较好的泛化能力;在所有被测的模型中,KNN 虽然检测准确率最低,但在面对对抗性攻击时表现出了较好的鲁棒性。

Wu 等^[46]提出了一种基于强化学习的对抗性僵尸网络流量生成框架来欺骗黑盒检测模型。攻击框架应用了 Q-学习算法,会不断地向目标模型提供输入并查询布尔反馈,当反馈说明本次欺骗成功时会对智能体进行奖励,Agent 的每个动作都是一组提前定义好的对流量进行修改的操作。为了方便、简洁地表示样本的状态,在输入攻击框架之前,还使用自动编码器对数据进行了预处理。选择 CTU-13 数据集^[47],训练了两种不同的模型进行实验:卷积神经网络和决策树。利用该框架生成的对抗性僵尸网络流量在 CNN 上获得了 35%~50% 的规避率,但在决策树上的平均规避率却低于 10%。其推测这可能是由于 CNN 模型在预处理环节需要先把流量数据转换成图像数据,添加到样本中的扰动可以反映像素的变化,这在检测过程中将更加明显。

Sharon 等^[48]通过对数据包间的延迟进行重塑,在不改变数据包内容的前提下首次实现了针对黑盒 IDS 的端到端规避攻击。利用长短期记忆网络选择 CIC-IDS2017 数据集和 Kitsune 数据集^[4]生成对抗样本,对 Autoencoder, KitNET 和 Isolation Forest 3 种检测器发起攻击。LSTM 的输入是前一个数据包时间戳的简短历史记录,以及前一个数据包和当前数据包的大小。在端到端的场景中,当数据包到达目标网络时检测器会逐包处理它们并做出响应决策,经过训练的 LSTM 可以根据目标网络的延迟和响应实时地调整每个数据包的发送时间戳,以规避目标检测器的检测。通过利用数据包的时间戳属性对攻击流量进行重塑,在端到端场景中以 99.99% 的成功率实现了检测逃逸。最后,还提出了一种基于对抗性训练的防御策略,用于缓解来自此类攻击的威胁。这种对流量的时序特征进行扰动以生成对抗性网络流量的过程必须非常谨慎,否则在某些情况下可能会违反网络流量的物理传输规则。

4.5 基于迁移的生成方法

基于迁移的生成方法适用于黑盒攻击。攻击者可以通过获取模型的少量输入输出训练出替代模型,随后对替代模型实施白盒攻击,利用迁移性在不掌握目标模型相关信息的前提下生成对抗样本,该方法可以被认为是基于决策的生成

方法的一种变体形式。

Usama 等^[49]提出了一种针对匿名网络流量分类的黑盒对抗性攻击。攻击者可以向目标模型提供输入并获得响应的标签,并将每个查询及其响应标签作为合成数据对进行存储,然后根据交互信息训练出替代模型来模拟目标模型的决策边界,利用对抗样本的迁移性欺骗目标模型。选择 UNB-CIC 数据集^[50]中的 Tor 类别流量,针对深度神经网络和支持向量机通过训练替代的 DNN 模型发起攻击。在保持流量的可执行性与攻击性的前提下,从合成数据对中选择最具有辨识度的特征进行扰动,并对添加的扰动大小进行了限制。实验结果表明,在面对对抗性攻击时,DNN 和 SVM 的准确率、精确率、召回率和 F_1 值等性能指标都出现了明显降低的现象,证明了该方法的有效性。

Yang 等^[51]也利用 NSL-KDD 数据集对于深度神经网络在黑盒假设下应用了 3 种攻击策略以生成对抗样本:基于训练替代模型的攻击、基于 ZOO 的攻击和基于 GAN 的攻击。在第一个场景中,攻击者首先训练了一个与 DNN 有相似结构的替代模型,该模型由两个隐藏层构成,每层包含 128 个神经元,然后利用 C&W 攻击针对替代模型生成对抗样本;在第二个场景中,攻击者直接利用 ZOO 对 DNN 发起黑盒攻击;在第三个场景中,攻击者基于 W-GAN 框架生成对抗样本。针对不同的实验场景进行了评估:在第一个场景中,发现基于训练替代模型的攻击破坏性似乎最小, F_1 值从初始的 0.898 变为 0.687 只降低了 24%,猜测原因可能是替代的模型还不能充分地表达目标模型的功能和结构或者对抗样本的迁移性不太稳定;在第二个场景中,发现基于 ZOO 的攻击效果最好, F_1 值降为了 0.273,但由于查询和估计梯度需要较大的计算成本,因此认为 ZOO 并不太适用于现实场景;在第三个场景中,发现基于 GAN 的攻击也取得了良好的效果, F_1 值降为 0.350,但在训练的过程中存在着收敛失败和模型崩溃等问题,还需要进一步的研究。

5 相关问题的讨论

5.1 生成方法效能问题

在不同的应用场景下,攻击者通常会选择不同的方法生成对抗性网络流量,不同的生成方法之间没有绝对的优劣之分,针对特定的目标模型需要根据实际的应用场景选择合适的对抗性网络流量生成方法发起攻击。我们根据攻击者对目标模型的了解程度,对生成方法的效能问题进行了分类讨论。

5.1.1 白盒攻击

在白盒攻击假设下,攻击者完全掌握与目标模型相关的所有信息,包括参数、结构、激活函数等。攻击者可以选择基于梯度的生成方法和基于优化的生成方法。现有的研究工作表明,在图像分类领域取得成功的攻击策略并不一定都适用于入侵检测领域。

FGSM、BIM 和 PGD 会使得原始流量数据的全部特征都发生微小的改变,无法保证流量修改后的可执行性与攻击性,JSMA、DeepFool 和 C&W 攻击是更符合实际的攻击策略。

通常情况下,JSAM 的欺骗效果最好但时间成本最高,DeepFool 的欺骗效果与时间成本居中,C&W 攻击的欺骗效果相对较差但时间成本最低。在面对大型数据集或者防御策略时,相比 JSMA 和 DeepFool,C&W 攻击会表现出更好的效能。

在此假设下,如果从欺骗效果和时间开销的角度进行选择,DeepFool 方法更为合适;如果从鲁棒性的角度进行选择,则 C&W 攻击会更为合适。

5.1.2 黑盒攻击

在黑盒攻击假设下,攻击者对目标模型的了解非常有限,只能提供输入数据并查询模型的输出。攻击者可以选择基于优化的生成方法、基于 GAN 的生成方法、基于决策的生成方法和基于迁移的生成方法。

基于优化的生成方法适用范围最广,泛化能力和攻击效果都较好,但是在面对大型数据集时由于搜索空间过大,效能不高;基于 GAN 的生成方法可以学习到高维复杂的数据分布而不需要依赖任何先验假设,但是每次只能为某一种数据类别生成对抗性网络流量,在博弈的过程中还可能会存在收敛失败和模型崩溃等问题;基于决策的生成方法和基于迁移的生成方法可以在不掌握模型参数、结构等信息的基础上实施黑盒攻击,但通常需要大量的迭代次数,存在时间开销较大的问题。

在此假设下,如果面对的是中小型数据集,应该优先选择基于优化的生成方法;如果面对的是大型数据集,且需要对抗的流量类别较少时,应该优先选择基于 GAN 的生成方法;如果面对的是大型数据集,且需要对抗的流量类别较多时,应该优先选择基于决策的生成方法或者基于迁移的生成方法。

5.2 特征扰动添加问题

为了保持对抗性网络流量的可执行性与攻击性不变,必须将扰动添加的范围约束在由非功能性特征组成的可行域之内。通过我们的调查发现,现有的研究工作主要采取两类方法进行:1)对预处理后提取的特征进行修改;2)直接对网络流中的数据包进行修改。但是,流量数据从输入空间到特征空间的映射过程是不可逆也不可微分的,第一类方法虽然可以实现检测逃逸,但如何将其还原到流量空间需要额外的算法支持和计算开销。因此,我们重点研究第二类方法,探讨在实际的操作过程中,应该被修改的数据包以及如何修改的问题。

网络流中的前几个数据包处于网络连接的协商阶段,其顺序通常是预先定义的,且不同应用的数据包大小是可明确区分的,含有丰富的特征信息。Hartl 等^[32]和 Sadeghzadeh 等^[35]的研究都已经表明:基于机器学习的流量检测结果主要受网络流起始部分数据包的影响,在后续的时间步骤中修改特征基本不会再改变其置信度。因此,应该将网络流起始部分的数据包作为修改对象。

恶意流量的非功能性特征可以分为内容特征与时序特征两类。在确定了修改对象之后,具体的数据包修改操作可以根据特定流量类别的非功能性特征被分类讨论。

(1)内容特征:主要指数据包中携带的文本信息,有效载荷中可能含有一些特殊的字节序列,是 DoS 类和 Probe 类等恶意流量的非功能性特征。攻击者可以尝试向流量负载中添加一些噪声数据或在网络流起始部分注入几个精心构造的虚假数据包,虚假数据包的载荷部分只是一些随机的字节序列。

(2)时序特征:主要指数据包中携带的统计信息,如数据包的平均长度、传输速率和传输方向等,是 U2R 类和 R2L 类等恶意流量的非功能性特征。攻击者可以对数据包的载荷大小、发送速率和传输时间间隔进行调整,在条件允许的情况下还可以尝试构造虚假的双向流量。但是,在对时序特征进行

扰动的过程中不能违反网络流量的物理传输规则。

5.3 网络流量数据问题

手动标记实际网络流量需要大量的时间成本,并且隐私问题也阻碍了实际网络流量的共享。因此,现有的研究工作主要是利用一些基准数据集进行对抗性网络流量的生成与应用。网络流量数据集可以分为3类:1)基于包的数据集,流量数据中包含完整的有效载荷信息,主要以PCAP格式进行

存储;2)基于流的数据集,流量数据中只包含网络连接的元数据,主要以会话的形式进行存储;3)混合数据集,流量数据中既包含网络连接的元数据又包含部分载荷信息。

表1对本文中提及的基准数据集进行了概述。表3则从敌手知识、敌手目标、攻击技术、检测器/模型、实验数据集和流量研究粒度6个维度对近年来对抗性网络流量的相关工作进行了梳理和归纳。

表1 文献综述中使用的数据集总结

Table 1 Summary of datasets used in literature review

Ref.	Datasets	Description
[41]	KDDCup99	-Dataset simulates 4 majorcyber attacks;DoS,Prob,Remote to Local and User to Root attacks.
[25]	NSL-KDD	-Intrusion dataset withDoS,Probe,User to Root and Remote to User attacks.
[31]	Moore	-Dataset contains 377526 network samples of 12 application categories,with a total of 248 attributes.
[45]	DARPA2009	-Dataset comprises DoS and worms which are parametrized to exhibit various propagation characteristics.
[47]	CTU-13	-Labeled dataset generated at CTU university containing botnet and benign traffic.
[50]	UNB-CIC	-Dataset consists of Tor network traffic provided by 8 different applications.
[34]	UNSW-NB15	-Dataset contains 49 features with class labels and 9 types of attacks(Fuzzers,Backdoors,DoS,Shellcode...)
[36]	ISCXVPN2016	-Dataset consists of 6 types of network traffic in normal environment and VPN environment.
[33]	CIC-IDS2017	-Intrusion dataset numerous types of attacks(DoS,PortScan,Infiltration,Botnet,XSS...)
[4]	Kitsune	-Dataset contains 8 network attacks,along with the benign and malicious labels for the network traffic packets.
[37]	Alexa-200	-Dataset randomly selected 200 different domain names from Alexa’s top 1000 rankings.
[27]	BoT-IoT	-Dataset consists over 72 million records of network activity in a simulated IoT environment.

表2 对抗性网络流量研究工作总结

Table 2 Summary of research work on adversarial network traffic

Paper	Year	Adversarial Knowledge	Adversarial Goals	Attack Techniques	Detectors/Models	Datasets	Traffic Granularity
[19]	2017	White-box	Non-targeted	FGSM,JSMA	DT,RF,LSVM, Voting Method	NSL-KDD	Flow-level
[26]	2018	White-box	Non-targeted	FGSM,JSMA, JSMA-RE	MLP,CNN	Port Scan Attack Dataset Based on SDN	Flow-level
[22]	2019	White-box	Non-targeted	FGSM,BIM,PGD	FNN,SNN	BoT-IoT	Flow-level
[28]	2021	White-box	Targeted & Non-targeted	FGSM	ANN	BoT-IoT	Flow-level
[29]	2018	White-box	Targeted & Non-targeted	FGSM,JSMA, DeepFool,C&W	MLP	NSL-KDD	Flow-level
[30]	2020	White-box	Non-targeted	FGSM,C&W,DeepFool	LeNet-5	Moore	Flow-level
[32]	2020	White-box	Non-targeted	FGSM,PGD,C&W	RNN	CIC-IDS2017, UNSW-NB15	Flow-level & Packet-level
[35]	2021	Black-box	Non-targeted	UAP	CNN	ISCXVPN2016	Flow-level
[21]	2019	Black-box	Non-targeted	Genetic Algorithm	Profile HMM	Alexa-200	Flow-level
[38]	2019	Black-box	Non-targeted	GAN	AlexNet	UNSW-NB15	Flow-level
[20]	2018	Black-box	Non-targeted	IDSGAN	SVM,NB,MLP,LR, DT,RF,KNN	NSL-KDD	Flow-level
[40]	2019	Black-box	Non-targeted	DoS-WGAN	CNN	KDDCup99	Flow-level
[42]	2019	Black-box	Non-targeted	GAN	DNN,LR,SVM,KNN, NB,RF,DT,GB	KDDCup99	Flow-level & Packet-level
[43]	2019	Black-box	Non-targeted	Decision Feedback	Kitsune,DAGMM,BiGAN	CIC-IDS2017,	Packet-level
[44]	2020	Black-box	Non-targeted	Decision Feedback	LR,RF,SVN,KNN	DARPA2009	Packet-level
[46]	2019	Black-box	Non-targeted	Q-learning	CNN,DT	CTU-13	Flow-level
[48]	2021	Black-box	Non-targeted	Decision Feedback	KitNET,Autoencoder, Isolation Forest	CIC-IDS2017,Kitsune	Packet-level
[49]	2019	Black-box	Non-targeted	Substitute Model	DNN,SVM	UNB-CIC	Flow-level
[51]	2018	Black-box	Non-targeted	ZOO,GAN, Substitute Model	DNN	NSL-KDD	Flow-level

6 未来研究工作

通过对抗性网络流量在入侵检测领域相关工作中的探讨,发现以下 4 个现象:

(1)在模型训练阶段,当输入数据的特征规范化时,基于机器学习的 IDS 虽然具有更好的检测性能,但是可能也更容易遭受对抗样本的攻击;

(2)在模型攻击阶段,现有的研究工作主要是利用不同的网络流量数据集对目标模型进行离线测试,很少在具有实时流量的环境中实施验证;

(3)不同的特征在被攻击者选择干扰的比率方面具有不同程度的显著性,恶意流量的非功能性特征更容易造成模型的脆弱性,在检测和防御工作中,它们应该得到更多的关注和更好的保护;

(4)虽然针对抗性网络流量的防御策略还未被充分探索,但多项研究工作都表明对抗性训练可以有效缓解来自对抗性网络流量的威胁。

基于上述现象,我们认为在未来工作中,以下 3 个方向值得重点研究:

(1)利用丰富的数据集资源,在具有实时流量的环境中采用直接对数据包进行修改的方法,进一步研究对抗性网络流量的生成与应用,使研究结论具有更现实的指导意义;

(2)构建一个完善且具有权威性的体系来评估在所有场景下机器学习模型鲁棒性的问题,并将模型鲁棒性作为基于机器学习的 IDS 在开发测试阶段一项重要的性能度量标准;

(3)进一步研究对抗性网络流量的防御策略,可以考虑将拟态防御等新的技术与机器学习相结合以增强算法的稳健性,引入深度学习模型可解释性方面的研究对检测模型进行调整与优化等。

结束语 本文探讨了将对抗性网络流量应用于入侵检测领域的一些研究工作。首先介绍了对抗性网络流量的基本概念、威胁模型和评价指标;然后对图像分类领域中用于生成对抗样本的攻击策略进行介绍;在此基础上,按照其生成方式与原理的不同分为 5 类,介绍了对抗性网络流量的生成方法以及其在入侵检测领域的应用;通过对相关问题的梳理总结,提出了未来的研究方向。

参 考 文 献

- [1] LUNT TF. A survey of intrusion detection techniques[J]. Computers & Security, 1993, 12(4): 405-418.
- [2] CHALAPATHY R, CHAWLA S. Deep Learning for Anomaly Detection: A Survey[J]. arXiv:1901.03407, 2019.
- [3] KWON D, KIM H, KIM J, et al. A survey of deep learning-based network anomaly detection[J]. Cluster Computing, 2019, 22(1): 949-961.
- [4] MIRSKY Y, DOITSHMAN T, ELOVICI Y, et al. Kitsune: An Ensemble of Autoencoders for Online Network Intrusion Detection[C] // Network and Distributed System Security Symposium, 2018.
- [5] MEGHDOURI F, ZSEBY T, IGLESIAS F. Analysis of Lightweight Feature Vectors for Attack Detection in Network Traffic[J]. Applied Sciences, 2018, 8(11).
- [6] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing

- properties of neural networks[J]. arXiv:1312.6199, 2013.
- [7] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and Harnessing Adversarial Examples[J]. arXiv:1412.6572, 2014.
- [8] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial Machine Learning at Scale[J]. arXiv, 2016.
- [9] PAPERNOT N, MCDANIEL P, JHA S, et al. The Limitations of Deep Learning in Adversarial Settings[C] // 2016 IEEE European Symposium on Security and Privacy (EuroS&P). 2015.
- [10] 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. 2016: 2574-2582.
- [11] CARLINI N, WAGNER D. Towards Evaluating the Robustness of Neural Networks[C] // 2017 IEEE Symposium on Security and Privacy (SP). 2017.
- [12] CHEN P Y, ZHANG H, SHARMA Y, et al. ZOO: Zeroth Order Optimization based Black-box Attacks to Deep Neural Networks without Training Substitute Models[C] // Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. 2017: 15-26.
- [13] MOOSAVI-DEZFOOLI S M, FAWZI A, FAWZI O, et al. Universal adversarial perturbations[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1765-1773.
- [14] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative Adversarial Nets[J]. MIT Press, 2014.
- [15] PAPERNOT N, MCDANIEL P, GOODFELLOW I, et al. Practical Black-Box Attacks against Deep Learning Systems using Adversarial Examples[J]. arXiv:1602.02697, 2016.
- [16] FAWZI A, FAWZI O, FROSSARD P. Fundamental limits on adversarial robustness[C] // Proceedings of ICML, Workshop on Deep Learning. 2015.
- [17] TABACOF P, VALLE E. Exploring the Space of Adversarial Images[C] // 2016 International Joint Conference on Neural Networks (IJCNN). 2016.
- [18] AKHTAR N, MIAN A. Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey[J]. IEEE Access, 2018, 6: 14410-14430.
- [19] RIGAKI M. Adversarial Deep Learning Against Intrusion Detection Classifiers[C] // 017 NATO IST-152 Workshop on Intelligent Autonomous Agents for CyberDefence and Resilience, IST-152 2017. 2017.
- [20] LIN Z, SHI Y, XUE Z. IDSGAN: Generative Adversarial Networks for Attack Generation against Intrusion Detection[J]. arXiv:1809.02077, 2018.
- [21] LIU X, ZHUO Z, DU X, et al. Adversarial attacks against profile HMM website fingerprinting detection model[J]. Cognitive Systems Research, 2019, 54(MAY): 83-89.
- [22] IBITOYE O, SHAFIQ O, MATRAWY A. Analyzing Adversarial Attacks against Deep Learning for Intrusion Detection in IoT Networks[C] // 2019 IEEE Global Communications Conference (GLOBECOM). 2019.
- [23] TRAMÉR F, KURAKIN A, PAPERNOT N, et al. Ensemble Adversarial Training: Attacks and Defenses[J]. arXiv: 1705.07204, 2017.
- [24] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards Deep

- Learning Models Resistant to Adversarial Attacks[J],2017.
- [25] TAVALI M,BAGHERI E,LU W,et al. A detailed analysis of the KDD CUP 99 data set[C]// IEEE International Conference on Computational Intelligence for Security & Defense Applications, IEEE,2009.
- [26] HUANG C H,LEE T H,CHANG L H,et al. Adversarial Attacks on SDN-Based Deep Learning IDS System[C]// Springer, Singapore,2018.
- [27] KORONIOTIS N,MOUSTAFA N,SITNIKOVA E,et al. Towards the Development of Realistic Botnet Dataset in the Internet of Things for Network Forensic Analytics; Bot-IoT Dataset [J]. Future Generation Computer Systems,2019,100:779-796.
- [28] PAPADOPOULOS P,ESSEN O,PITROPAKIS N,et al. Launching Adversarial Attacks against Network Intrusion Detection Systems forIoT[J]. Journal of Cybersecurity and Privacy,2021,1(2):252-273.
- [29] WANG Z. Deep Learning-Based Intrusion Detection With Adversaries[J]. IEEE Access,2018,6:38367-38384.
- [30] HU Y J,GUO Y B,MA J,et al. Method to generate cyber deception traffic based on adversarial sample[J]. Journal on Communications,2020,41(9):59-70.
- [31] ANDREW W,MOORE D. Discriminators for use in flow-based classification[R]. UK:Queen Mary University of London,2005.
- [32] HARTL A,BACHL M,FABINI J,et al. Explainability and Adversarial Robustness for RNNs[C]// 2020 IEEE Sixth International Conference on Big Data Computing Service and Applications(BigDataService). 2020.
- [33] SHARAFALDIN I,LASHKARI A H,GHORBANI A A. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization[C]// International Conference on Information Systems Security & Privacy. 2018.
- [34] MOUSTAFA N,SLAY J. UNSW-NB15:a comprehensive data set for network intrusion detection systems(UNSW-NB15 network data set)[C]// Military Communications and Information Systems Conference(MilCIS). 2015.
- [35] SADEGHZADEH A M,SHIRAVI S,JALILI R. Adversarial Network Traffic: Towards Evaluating the Robustness of Deep-Learning-Based Network Traffic Classification [J]. IEEE Transactions on Network and Service Management,2021,18(2): 1962-1976.
- [36] LASHKARI A H,DRAPER-GIL G,MAMUN M,et al. Characterization of Encrypted and VPN Traffic Using Time-Related Features[C]// The International Conference on Information Systems Security and Privacy(ICISSP). 2016.
- [37] ZHUO Z,ZHANG Y,ZHANG Z L,et al. Website fingerprinting attack on anonymity networks based on profile hidden markov model[J]. IEEE Transactions on Information Forensics & Security,2017,13(5):1081-1095.
- [38] PAN Y M,LIN J J. Malicious Network Stream Generation and Verification Based on Generative Adversarial Networks[J]. Journal of East China University of Science and Technology, 2019,45(2):165-171.
- [39] ARJOVSKY M,CHINTALA S,BOTTOU L. Wasserstein generative adversarial networks[C]// International Conference on Machine Learning, PMLR,2017:214-223.
- [40] YAN Q,WANG M,HUANG W,et al. Automatically synthesizing DoS attack traces using generative adversarial networks [J]. International Journal of Machine Learning and Cybernetics, 2019,10(12):3387-3396.
- [41] ZHANG X Y,ZENG H S,JIA L. Research of intrusion detection system dataset-KDD CUP99[J]. Computer Engineering and Design,2010,31(22).
- [42] USAMA M,ASIM M,LATIF S,et al. Generative Adversarial Networks For Launching and Thwarting Adversarial Attacks on Network Intrusion Detection Systems[C]// 2019 15th International Wireless Communications and Mobile Computing Conference(IWCMC). 2019.
- [43] HASHEMI M J,CUSACK G,KELLER E. Towards Evaluation of NIDSs in Adversarial Setting[C]// Proceedings of the 3rd ACMCoNEXT Workshop on Big Data, Machine Learning and Artificial Intelligence for Data Communication Networks: Association for Computing Machinery. 2019.
- [44] AIKEN J,SCOTT-HAYWARD S. Investigating Adversarial Attacks against Network Intrusion Detection Systems in SDNs [C]// 2019 IEEE Conference on Network Function Virtualization and Software Defined Networks(NFV-SDN). 2020.
- [45] BAHROLOLUM M,SALAH E,KHALEGHI M. Anomaly Intrusion Detection Design using Hybrid of unsupervised and supervised Neural Network[J]. International Journal of Computer Networks & Communications(IJCNC),2009,1(2):26-33.
- [46] WU D,FANG B,WANG J,et al. Evading Machine Learning Botnet Detection Models via Deep Reinforcement Learning [C]// 2019 IEEE International Conference on Communications (ICC 2019). 2019.
- [47] GARCIA S,GRILL M,STIBOREK J,et al. An empirical comparison of botnet detection methods[J]. Computers & Security, 2014,45(Sep.):100-123.
- [48] SHARON Y,BEREND D,LIU Y,et al. TANTRA:Timing-Based Adversarial Network Traffic Reshaping Attack[J]. arXiv:2103.06297,2021.
- [49] USAMA M,QAYYUM A,QADIR J,et al. Black-box Adversarial Machine Learning Attack on Network Traffic Classification [C]// 2019 15th International Wireless Communications and Mobile Computing Conference(IWCMC). 2019.
- [50] LASHKARI A H,GIL G D,MAMUN M,et al. Characterization of Tor Traffic using Time based Features[C]// International Conference on Information Systems Security & Privacy. 2017.
- [51] YANG K,LIU J,CHI Z,et al. Adversarial Examples Against the Deep Learning Based Network Intrusion Detection Systems [C]// 2018 IEEE Military Communications Conference (MILCOM 2018). 2018.



WANG Jue, born in 1998, postgraduate. His main research interests include cybersecurity and intrusion detection.



ZHU Yue-fei, born in 1962, Ph.D, professor, Ph.D supervisor. His main research interests include cryptography, intrusion detection and information security.