对抗机器学习在网络入侵检测领域的应用

刘奇旭[1,2]，王君楠[1,2]，尹捷[1]，陈艳辉[1,2]，刘嘉熹[1,2]
（1. 中国科学院信息工程研究所，北京 100093；2. 中国科学院大学网络空间安全学院，北京 100049）

摘 要：近年来，机器学习技术逐渐成为主流网络入侵检测方案。然而机器学习模型固有的安全脆弱性，使其难以抵抗对抗攻击，即通过在输入中施加细微扰动而使模型得出错误结果。对抗机器学习已经在图像识别领域进行了广泛的研究，在具有高对抗性的入侵检测领域中，对抗机器学习将使网络安全面临更严峻的安全威胁。为应对此类威胁，从攻击、防御 2 个角度，系统分析并整理了将对抗机器学习技术应用于入侵检测场景的最新工作成果。首先，揭示了在入侵检测领域应用对抗机器学习技术所具有的独特约束和挑战；其次，根据对抗攻击阶段提出了一个多维分类法，并以此为依据对比和整理了现有研究成果；最后，在总结应用现状的基础上，讨论未来的发展方向。

关键词：入侵检测；恶意流量；对抗攻击；对抗防御
中图分类号：TN92
文献标识码：A
DOI: 10.11959/j.issn.1000−436x.2021193

Application of adversarial machine learning in network intrusion detection

LIU Qixu[1,2], WANG Junnan[1,2], YIN Jie[1], CHEN Yanhui[1,2], LIU Jiaxi[1,2]
1. Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
2. School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: In recent years, machine learning (ML) has become the mainstream network intrusion detection system(NIDS). However, the inherent vulnerabilities of machine learning make it difficult to resist adversarial attacks, which can mislead the models by adding subtle perturbations to the input sample. Adversarial machine learning (AML) has been extensively studied in image recognition. In the field of intrusion detection, which is inherently highly antagonistic, it may directly make ML-based detectors unavailable and cause significant property damage. To deal with such threats, the latest work of applying AML technology was systematically investigated in NIDS from two perspectives: attack and defense. First, the unique constraints and challenges were revealed when applying AML technology in the NIDS field; secondly, a multi-dimensional taxonomy was proposed according to the adversarial attack stage, and current work was compared and summarized on this basis; finally, the future research directions was discussed.

Keywords: intrusion detection, malicious traffic, adversarial attack, adversarial defense

---

收稿日期：2021−07−20；**修回日期**：2021−09−15
通信作者：尹捷，yinjie@iie.ac.cn
基金项目：中国科学院青年创新促进会基金资助项目（No.2019163）；国家自然科学基金资助项目（No.61902396）；中国科学院战略性先导科技专项基金资助项目（No.XDC02040100）；中国科学院网络测评技术重点实验室和网络安全防护技术北京市重点实验室基金资助项目

Foundation Items: The Youth Innovation Promotion Association CAS (No.2019163), The National Natural Science Foundation of China (No.61902396), The Strategic Priority Research Program of Chinese Academy of Sciences (No.XDC02040100), The Key Laboratory of Network Assessment Technology at Chinese Academy of Sciences and Beijing Key Laboratory of Network Security and Protection Technology

1 引言

网络入侵检测系统（NIDS, network intrusion detection system）是一种主动安全防护技术，通过实时监视网络流量，并与已知的攻击库进行匹配[1]，从而快速感知被监控网络中违背安全策略、危及系统安全的行为，是保障网络安全的重要手段。

黑客攻击手段的不断升级及网络数据的大规模增长，让传统基于规则的检测方法难以为继，系统占用资源过多、对未知攻击检测能力差、需要人工干预等缺点日益突出[2]。随着人工智能的蓬勃发展，基于机器学习（ML, machine learning）的异常检测逐渐成为主流 NIDS 方案[3]。ML 算法的高速运算能力能够有效应对大规模网络流量，且 ML 模型应对分类问题具有天然优势，其出色的泛化能力可使 NIDS 具备一定检测未知恶意流量的能力。

机器学习作为一个复杂的计算系统，同样面临着安全性考验。2014 年，Szegedy 等[4]首次提出了对抗样本概念，发现可以通过在原始样本中加入精心构造的微小扰动从而误导 ML 模型，并将这种形式的攻击命名为对抗攻击。随后 Papernot 等[5]发现对抗样本具有在不同 ML 模型之间迁移的特性。这一发现揭露了 ML 技术在安全方面的极大缺陷，从而使人们更加谨慎看待 ML 算法在其他领域的应用。2017 年，Kurakin 等[6]用手机相机拍摄生成的洗衣机对抗样本照片，成功误导了 TensorFlow Camera Demo 应用。2018 年，Alzantot[7]通过变动少量词汇成功攻击了情感分析和文本蕴含模型。2019 年，Qin 等[8]构造了人耳无法辨别的音频对抗样本。在一些安全敏感的领域，如医疗检测[9]、自动驾驶[10]和人脸识别[11]等，对抗攻击将带来更加严峻的安全威胁，直接影响人们的人身、财产和隐私的安全。

入侵检测原本就是具有高对抗性的网络攻防对抗领域，对抗攻击的出现为恶意网络攻击者提供了强大武器，给安全防御工作带来极大挑战。因此对抗攻击在入侵流量检测与绕过这一领域的应用引起了安全研究者的广泛关注[12-17]。

本文首先结合 NIDS 领域特有的流量对象和场景需求，提出了将对抗机器学习（AML, adversarial machine learning）技术应用于 NIDS 领域的独特约束和挑战。其次，结合上述约束和威胁模型，本文从攻防 2 个视角提出适用于 NIDS 领域的对抗攻击和对抗防御多维分类法，从多个角度总结和对比现有工作。最后，基于对已有工作的总结分析，本文对当前研究的不足和未来发展方向进行探讨和展望。

2 基础知识

2.1 基本概念及形式化

机器学习的对抗样本问题引起了研究者的极大关注，并提出了一系列的对抗攻击和对抗防御的方法，这一领域即对抗机器学习。

如图 1 所示，对抗样本是指在测试样本上添加精心构造的扰动而生成的新的输入样本。对抗样本与原始测试样本在人为观察上无明显差异，但会使 ML 模型产生与原始样本完全不同的预测结果。

对抗性扰动则是根据 AML 算法，通过最大限度提高预测误差而获得的难以察觉的非随机噪声。

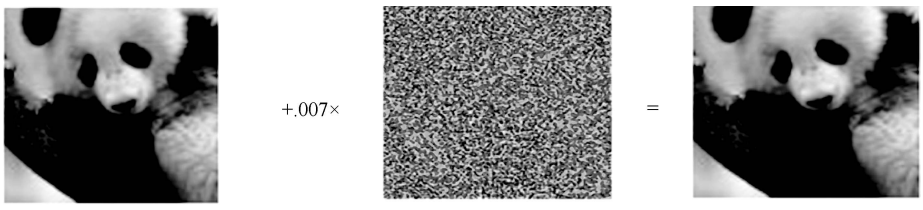
给定经过训练的 ML 分类器 $f:R^m\to\text{Label}$ 和原始输入数据样本 $x\in[a,b]^m$，生成对抗样本 $x'=x+\eta$（η 为对抗性扰动）的过程通常可以形式化为一个边界约束的优化问题[4]，即

$$\begin{aligned}&\min\|x'-x\|\\&\text{s.t.}f(x')=y',f(x)=y,y\neq y',\\&x'\in[a,b]^m,y,y'\in\text{Label}\end{aligned}\tag{1}$$

其中，y' 表示 x' 的目标标签；$\|\bullet\|$ 表示 2 个样本的距离；边界约束 $x'\in[a,b]^m$ 意味着必须在有效边界内生成对抗样本；$\eta=x'-x$ 表示添加在原始样本 x 上的扰动。该优化问题的目标就是在最小化扰动的同时，使模型将添加了扰动的输入数据错误分类。

2.2 经典对抗机器学习算法

Goodfellow 等[18]提出了快速梯度符号方法



原始样本x
被ML模型识别为熊猫

微小扰动η

对抗样本$x'=x+\eta$
被ML模型识别为长臂猿

图 1 对抗样本

（FGSM, fast gradient signed method），在 L∞ 范数限制下将样本朝着梯度上升方向推动，快速增加损失函数，从而改变分类结果。FGSM 计算成本低、生成速度快，但攻击能力较弱，适用于对攻击效率有较高要求的应用场景。

Papernot 等[19]提出基于雅可比矩阵的显著图攻击方法（JSMA, jacobian-based saliency map attack）。JSMA 通过计算正向导数来评估模型对每个输入特征的敏感度，进而定位最易导致模型输出发生重大变化的像素点来欺骗模型。JSMA 使用 L0 范数来限制扰动大小，对原始输入修改较少，同时由于 JSMA 采用正向传播计算显著点，计算过程相对简单。

Moosavi-Dezfooli 等[20]提出了能够生成最小化 L2 范数的对抗性扰动的方法——DeepFool。DeepFool 逐渐将位于分类边界一侧的样本推向另一侧，直到分类错误。这种方法对原始输入的改动相对较少，同时生成的对抗样本具有较好的攻击效果，但计算量也相对较高。

Carlini 和 Wagner[21]提出的 C&W 攻击重新设计了多步迭代攻击中的损失函数，使其在对抗样本中有较小的值，但在原始样本中有较大的值，因此可通过最小化该损失函数得到对抗样本，是目前最先进的白盒攻击方案。

Moosavi-Dezfooli 等[22]进一步证明了跨越数据及网络架构的通用对抗扰动（UAP, universal adversarial perturbations）的存在。该方法对所有样本进行迭代版的 DeepFool 攻击，直到找到一个可以使大部分样本都分类错误的扰动 η。UAP 攻击不需要目标模型的任何信息，极大地降低了实施对抗攻击的门槛，危害性更大。

3　独特挑战与约束

在 NIDS 领域中，AML 扰动对象为恶意流量，这与在图片样本上进行修改有很大不同。本节结合 NIDS 领域的独特场景和对象，详细分析在 NIDS 领域应用 AML 算法所具有的独特挑战和约束。

3.1　独特的场景挑战

入侵检测与绕过领域具有高度对抗性，攻击者与防御者在技术更新迭代中不断对抗。在 NIDS 领域应用 AML 算法存在以下几点特殊挑战。

1) 应用场景的挑战。在真实的网络攻击场景中，攻击者往往拥有较少的检测模型知识，且无法通过不断访问 NIDS 来训练替代模型。因此基于梯度的 AML 算法和需要大量访问目标模型的攻击方法在 NIDS 场景中都无法直接应用。

2) 直接生成对抗流量的挑战。若想对 NIDS 产生真实威胁，攻击者需要产生能在网络传输的对抗流量，仅生成对抗特征向量的对抗攻击在 NIDS 领域难以适用。

3.2　独特的扰动约束

在生成图像对抗扰动时，往往以“微小”为基本约束，通过限制 L∞、L1、L2 范数使添加的扰动对人不可观察，从而避免被人类观察员发现。在 NIDA 领域，“微小”约束不再适用，取而代之的是恶意流量样本所带来的以下几点更严苛的约束条件。

1) 遵守网络协议规范的约束。为了保证被扰动后的网络流量仍能在现网传输，需要保证在添加扰动后仍符合网络协议规范。即各协议头部字段须保持正确含义，且符合各部分的值域范围。如 IP 层头部前 4 bit 代表版本号，值域范围是 $(4,6)$，在扰动后不应出现值域范围外的数值。

2) 维持恶意功能的约束。入侵流量往往承担着传递恶意信息、执行恶意功能的基本任务，为了保证在添加扰动后的恶意流量仍能实现原本恶意目的，添加的扰动不应对承担恶意功能的特征或区域产生影响。例如，随意更改恶意 payload 或将协议类型从 TCP 更改为 UDP，可能会导致恶意功能无法实现或是数据包无法传输。

3) 保证一致性的约束。如果扰动对象是恶意流量的特征向量，在施加扰动时还需要考虑各特征之间的相关关系。例如，特征向量同时包含持续时间 duration、数据包字节长度 Byte 和字节传输速率 bit/s，在施加扰动后仍应满足 bit / s = Byte/duration 的依赖关系，否则将无法据此生成真实恶意流量样本。

在 NIDS 领域应用 AML 算法所具有的独特的挑战和约束使很多原本适用于在计算机视觉领域的对抗攻击和防御方法无法适用，这些方法或因应用场景的严格限制和不透明而导致假设不成立，或因无法产生真实流量而不具备实际攻击价值，都很难直接应用于 NIDS 领域。为充分调研和分析相关工作成果，本文将结合本节提出的若干挑战和约束，构建适用于 NIDS 的 AML 分类法。

4　分类法

本文将基于机器学习的 NIDS 一般应用流程分类为模型训练和测试两大阶段，并以此为中心根据

不同攻击阶段提出了多维对抗攻击分类法，根据不同修正对象提出了与模型训练阶段各具体步骤相对应的对抗机器学习技术分类法，如图 2 所示。

4.1　对抗攻击分类法

考虑到在 NIDS 领域实施对抗攻击所具有独特约束和挑战，本文按照攻击的发展阶段提出了一种全新分类法。

如图 2 所示，作为一种测试时攻击，对抗攻击实施者将首先利用社会工程攻击或先验知识等侧信道信息充分分析攻防双方的形势地位以确立威胁模型，随后据此选择适当的对抗攻击方法，并生成对抗样本输入基于机器学习的 NIDS，从而绕过检测实现恶意目的。

威胁模型，是指本文通过对攻击者目标、能力以及知识的描述，建立一般性的威胁模型。

1) 攻击者知识，指攻击者对 ML 模型整体的了解程度。具体包括：全部或部分训练数据；样本的特征表示；模型所采取的机器学习算法和判别函数；训练后模型的具体参数；模型的判别结果[23]。根据信息掌握程度可以分为白盒、灰盒和黑盒攻击。

① 白盒攻击。假设攻击者对 NIDS 的训练数据、算法、模型及参数有完整的知识。白盒场景下防御者完全透明，在现实攻击场景中通常是不可行的。

② 灰盒攻击。假设对手可以访问不同程度的信息。对于灰盒攻击，对手没有模型创建者所拥有的确切知识，但有足够的信息攻击机器学习系统，导致机器学习系统失败。

③黑盒攻击。假设对手没有任何关于机器学习系统的先验知识，仅能通过访问模型获得二进制分类结果。真实场景中的黑盒攻击可能比理论上的黑盒模型更具限制性，例如在真实 NIDS 攻防场景中，攻击者往往仅具有有限的访问次数。

2) 攻击者能力。对抗攻击是测试时攻击，因此攻击者需具有操纵测试样本的能力。在 NIDS 场景下，攻击者往往是恶意软件的实际操控者，具有控制恶意流量的构造和发送，且根据要求调整恶意网络行为的较高能力权限。

3) 攻击者目的。对抗攻击的目的在于破坏机器学习系统的完整性，降低模型的可靠性。在 NIDS 二分类应用场景下，攻击者希望通过施加对抗性扰动，使基于机器学习的 NIDS 将对抗恶意流量样本识别为良性，从而绕过检测，增加攻击行为的隐蔽性。

攻击方法，是指根据攻击方法的基本原理进行分类。不同的攻击方法对攻击者的能力需求不同，产生的实际危害程度也会有所区别。

1) 基于梯度的攻击。攻击者掌握分类器的损失函数，通过获得目标模型对于特定输入的梯度信息，进行优化搜索或逆梯度构造来生成对抗样本。

2) 基于得分的攻击。攻击者利用目标模型的分

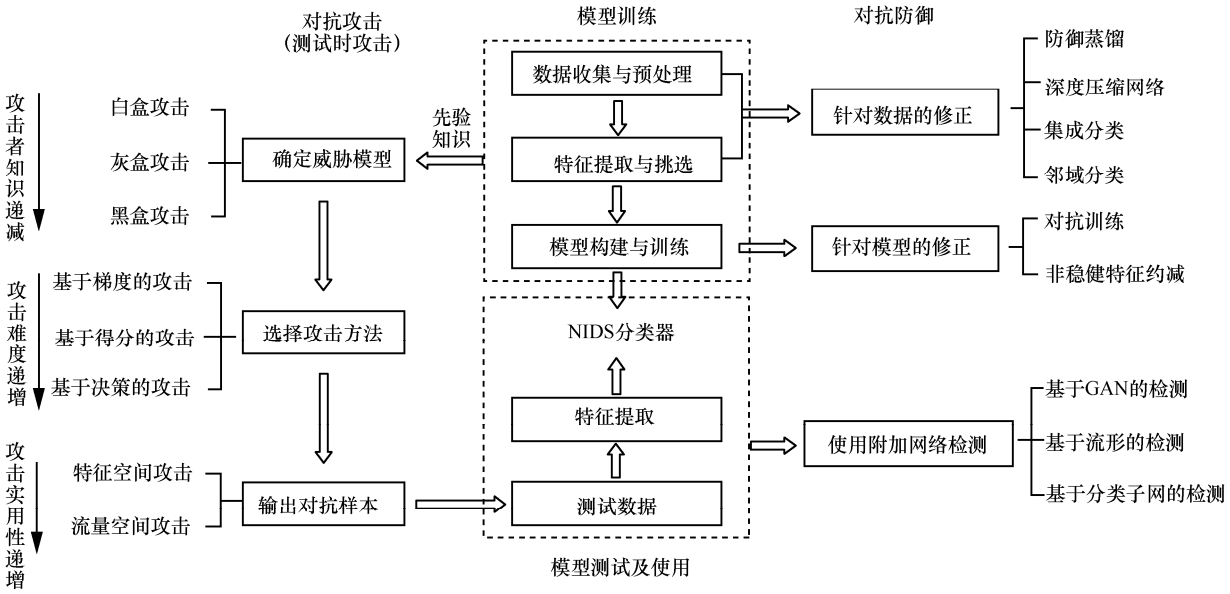


图 2　网络入侵检测领域对抗机器学习技术分类法

类置信度或 logits，即样本的概率向量，获得目标模型的细节信息，从而指导对抗样本的构造。

3) 基于决策的攻击。攻击者收集目标模型的二进制分类结果来训练替代模型，或利用生成模型学习对抗样本自身的分布，直接构造对抗样本的方法。

输出结果，是指根据对抗攻击最终生成的样本类型进行分类[24]。

在机器学习领域有问题空间和特征空间之分[25]。问题空间指样本实例的集合，特征空间则是用于表示样本实例的特征向量的集合。两者有时是相同的，更多时候特征空间是由问题空间通过某种映射生成的。根据对抗攻击方法的输出，可将其分为以下 2 种。

1) 特征空间攻击，指仅能生成具有对抗性的流量特征向量的对抗攻击方法。

2) 流量空间攻击，指能够生成可在网络中传输的流量样本的对抗攻击方法。

由于从流量样本到流量特征的映射通常是不可逆的，即在已知流量特征向量的情况下，很难通过逆映射运算得到真实流量样本。而且从特征到流量样本的映射往往并非双射，也就是说在给定特征向量 v 和对抗扰动 r，可能并不存在与 $v+r$ 对应的流量实例 x。因此在 NIDS 领域，特征空间和流量空间攻击的实际威胁程度有较大的差异。

4.2 对抗防御分类法

为降低对抗样本带来的安全威胁，增强 NIDS 模型的稳健性，安全防御人员需采取对抗防御措施来抵抗潜在的对抗攻击方法。如图 2 所示，根据作用阶段的不同，对抗防御可以分为在模型训练阶段针对数据的修正、针对模型的修正，以及在模型测试阶段使用附加网络检测对抗样本三类。

1) 针对数据的修正，是指通过在“数据收集与预处理”和“特征提取与挑选”阶段，通过改变模型的训练数据组成或改变模型的特征空间来间接地提高目标模型对对抗样本稳健性的一类防御方法。这类方法多数需要重新进行模型训练。

2) 针对模型的修正，是指通过直接修改目标分类模型实现对抗样本防御的一类方法。具体实现又可包括掩蔽梯度信息、增加模型正则化程度和采用更复杂的分类机制等方法。这类方法往往会增加模型复杂度，对模型的准确性也有一定影响。

3) 使用附加网络检测，是指在不改变目标分类模型的前提下，通过训练额外的 ML 模型检测对抗样本，从而排除对抗样本对目标分类器的影响。这类方法增加了训练附加网络的训练成本，而且可能会影响目标分类模型的效率和实用性。

5 对抗攻击

结合 NIDS 领域特有的约束，本文将根据第 4 节提到的分类法，从攻击者知识和输出结果 2 个角度总结分析当前对抗攻击在 NIDS 领域的具体实践。相关工作总结对比[26-44]如表 1 所示。

白盒&特征空间攻击。在白盒场景下的特征空间攻击方法利用传统基于梯度的 AML 算法，直接将网络流量特征向量当作图像样本施加无差别扰动，从而得到能够绕过基于 ML 模型的 NIDS 的、具有高对抗性的恶意流量特征向量。

文献[26]利用 FGSM、JSMA 方法针对多层感知机和由决策树、随机森林以及支持向量机构成的集成分类器发起对抗攻击。通过评估在 NSL-KDD 数据集上的分类器性能指标的下降（10%～27%）证明了基于机器学习的 NIDS 难以抵抗对抗攻击。文献[27]增加了 C&W 和 DeepFool 这 2 种攻击方法的评估。Ibitoye 等[28]则在 Bot-IoT 数据集[45]上评估了 FGSM、BIM、PGD 这 3 种攻击方法针对前馈神经网络和自归一化神经网络模型[29]的有效性。文献[46]则针对已发表的 IDS 模型 KitNet[47]实施了对抗攻击，并证明平均修改 1.38 个输入特征即可使 Mirai 恶意流量绕过检测。

白盒&流量空间攻击。为了生成可在网络中传输的攻击流量，在应用已有的基于梯度的攻击方法时，攻击者需要考虑在第 3 节中提到的几种特殊约束，从而限制对抗性扰动的影响范围。

文献[30]将扰动的特征限制在持续时间、总字节、总包数以及基本属性特征中，且仅采用 FGSM 生成的正值扰动。在这 2 项约束下，利用代理通过增加和时延数据包等操作实现了加密 C2 流量绕过检测，不但能够保证恶意功能和遵守网络协议规范，而且由于扰动的特征之间无依赖关系，该方法可以保证特征一致性。

文献[31]提出了一个融合特殊域约束的基于梯度的迭代对抗攻击框架，也提出了特征空间的特定域依赖和数学关系依赖，并针对性设计了可以保持特征内部一致性的更新函数。作者通过加包操作

表 1 **对抗攻击方法**

文献	年份	攻击目标	威胁模型	对抗攻击方法	输出	约束	影响的对象	数据集	评估指标
文献[26]	2017	MLP, 集成分类器	白盒	FGSM, JSMA	特征	No	流特征	NSL-KDD	acc, F1-score, AUC
文献[27]	2018	MLP	白盒	FGSM, JSMA, C&W, DeepFool	特征	No	流特征	NSL-KDD	acc, F1-score, P, R
文献[28]	2019	FNN, SNN	白盒	FGSM, BIM, PGD	特征	No	流特征	Bot-IoT	acc
文献[29]	2017	KitNet	白盒	FGSM, JSMA, C&W, ENM	特征	No	流特征	Mirai	FP, FN
文献[30]	2020	DNN	白盒	FGSM	流量	Yes	流特征	MTA	ER
文献[31]	2019	FFDNN, FFNN	白盒	PGD, C&W	流量	Yes	网络流, 流特征	CTU- 13	ASR, ROC 曲线
文献[32]	2021	1D-CNN	白盒	UAP	流量	Yes	数据包, 网络流, 流特征	ISCXVPN2016	R
文献[33]	2018	MLP	灰盒(特征)	C&W, ZOO, GAN	特征	No	流特征	NSL-KDD	acc, F1-score , P, R
文献[34]	2018	SVM, NB, MLP, LR, DT, RF, KNN	灰盒(特征)	WGAN	特征	Yes	流特征	NSL-KDD	DR, EIR
文献[35]	2020	GBDT	灰盒(特征)	Gen-AAL	特征	No	流特征	CIC-IDS2017	ASR
文献[36]	2019	LR, RF, SVM, KNN	灰盒(得分)	随机添加噪声	特征	Yes	流特征	DARPA	DR
文献[37]	2019	DNN, DT	灰盒(得分)	ATN：StarGAN	特征	Yes	流特征	Meek 流量	FPR, PR-AUC
文献[38]	2021	MLP, RF, GB, LR, LDA, QDA, BAG, 集成分类器	灰盒(特征)	遗传算法，粒子群算法，GAN	特征	Yes	流特征	NSL-KDD, UNSW-NB-15	ER
文献[39]	2019	DAGMM, IF, AE, AnoGAN, ALAD, DSVDD, OC-SVM	灰盒(特征)	球形局部子空间	流量	Yes	流特征	CIC-IDS2018	ASR
文献[40]	2018	Stratosphere IPS	黑盒	GAN	流量	Yes	流特征	CC 流量	DR
文献[41]	2019	KitNet, DAGMM, BiGAN	黑盒	迭代搜索	流量	Yes	数据包，流特征	CIC-IDS2017	TPR, FPR
文献[42]	2019	DT CNN	黑盒	Deep Q-learmng	流量	Yes	网络流	CTU-13	acc, ER
文献[43]	2021	MLP, DT, LR, SVM	黑盒	SeqGAN+RL	流量	Yes	数据包	CTU-13	MAPE, AFR, ASR
文献[44]	2021	AE, KitNet, IS	黑盒	LSTM	流量	Yes	网络流	CIC-IDS2017	DR

注：acc 表示准确度（accuracy）；P 表示精确度（precision）；R 表示召回率（recall）；TPR 表示真阳率；FPR 表示假阳率；EIR 表示规避增长率（evasion increase rate），即未检测到的敌对恶意流量实例相对于原始恶意流量实例的增长速度，EIR=1−(adversarial detection ate)/(original detection rate)；ER/ASR 表示规避率（evasion rate）/攻击成功率（attack success rate），即攻击流量被识别成正常的百分比，与模型的召回率线性相关。

改变持续时间、总包数、总字节数等特征，使基于 CTU 数据集[48]的检测模型的 AUC 从 0.98 降为 0.21。

文献[32]借鉴了 UAP 的思想，通过构造通用对抗扰动来绕过 1D-CNN 模型。针对网络流量分类的不同输入空间分别提出 3 种构造 UAP 的攻击方法。AdvPad 在原始数据包有效负载中注入 UAP；AdvPay 将 UAP 作为有效负载构建新数据包插入原始流中；AdvBurst 将具有 UAP 生成的统计特征的虚拟包注入原始流中。该方法基于代理实现了在数据包层面上的增量操作，不但保证了前述若干约束，而且能够使检测器的召回率下降 20%～70%。

灰盒&特征空间攻击。利用有限的知识构造能够绕过目标 NIDS 模型的对抗特征向量。

Yang 等[33]在 NSL-KDD 上针对 MLP 模型分别使用 C&W、ZOO[49]和生成式对抗网络（GAN, generative adversarial network）3 种方法生成对抗样本。作者提出，尽管 ZOO 的攻击效果更好，但它需要大量的查询来生成敌对的示例，很难适应真实的网络攻击场景。文献[34]同样利用 WGAN[50]构造对抗样本。作者假定攻击者了解目标模型的特征空

间，利用判别器模拟目标模型，同时仅保留对非功能性特征的扰动，从而弥补了文献[36]的不足。

Shu 等[35]提出了一种结合主动学习和 GAN 的对抗攻击方法。利用基于边缘采样的主动学习选择距离目标模型决策边界更近的样本来训练替代模型，从而减少目标模型的访问次数。实验表明，仅通过 25 次模型访问即可实现 98.86%的绕过率。

Aiken 等[36]通过仅向 3 种 SYN 泛洪的典型特征（数据包大小、数据包传输速率和上下行流量比率）随机添加噪声，基于目标模型反馈的分类置信度实验最佳扰动值。实验结果表明，同时扰动 3 种特征可将 RF、线性回归和 SVM 的准确率降为 0。

文献[37]提出了一种基于对抗性转换网络的对抗方法。作者观察到 Meek 流量和 HTTPS 流量在有效负载长度分布和数据包到达间隔分布 2 个特征上有显著差异，因此基于 StarGAN[51]融合多个损失函数，最小限度地修改统计特征，从而将 Tor 网络的流量隐藏在 HTTPS 连接中。实验证明本文提出的攻击方法可将平均 FPR 从 0.183 提高到 0.834。

以上工作通过仅扰动非功能性特征的方法来满足约束，文献[38]则进一步总结并解决了以下 3 种约束。1) 二进制特征：仅二进制翻转。2) 保持恶意功能：仅执行增量操作。3) 特征依赖：联动修改具有相关关系的特征。在这 3 种约束下，通过遗传算法、粒子群算法和 GAN 进行优化搜索找到最小扰动。在 NSL-KDD 和 UNSW-NB-15[52]数据集上的实验证明，可实现 92.6%的绕过成功率。

灰盒&流量空间攻击。考虑到社会工程攻击和恶意流量检测广泛存在的先验知识，攻击者往往能够获得目标模型的部分知识（如算法、特征空间等），从而更有针对性地设计对抗攻击方法。

Kuppa 等[39]在原始样本的球形局部子空间搜索生成对抗样本并利用流形近似算法[53]来减少查询次数。作者仅扰动非功能性特征，并使用 Scapy 不断更新数据包以维持特征之间的依赖关系，从而实现对抗特征向量到对抗流量样本的逆向构造。

黑盒&流量空间攻击。黑盒场景提出更加严格的限制，即攻击者仅能利用 NIDS 的二进制判别结果来构造能够绕过检测的恶意对抗流量。因此，黑盒场景下的流量空间攻击与真实场景下的 NIDS 绕过攻击场景最一致，也最具实用性和威胁性。

Rigaki 等[40]提出了一种基于 GAN 的流量空间攻击方法。作者利用 GAN 模拟 Facebook 流量的统计特征，并将获得的对抗性特征传递给恶意代码，以便其构造符合“要求”的恶意流量，从而使 Stratosphere IPS 无法区分恶意流量与 Facebook 流量。FlowGAN[54]不再局限于 Facebook 流量，而是可以模拟任何“正常”网络流量动态改变流量特征绕过审查。但这 2 个工作中，GAN 仍然仅负责生成对抗性特征。攻击者需要对恶意软件源代码进行复杂修改才能实现从对抗特征到对抗流量的转换。

Hashemi 等[41]采用基于决策的思想，借助有限的目标 NIDS 反馈，在数据包层次上或网络流层次上迭代修改原始输入样本，从而生成对抗样本。为了满足前述约束，作者限制扰动动作为数据包分裂、数据包时延和数据包注入 3 种操作。该方法的不足在于需要向目标 NIDS 发送大量的询问。

Wu 等[42]将对抗样本构造问题建模为序列决策问题，利用深度强化学习生成对抗样本。作者设计了一个包含 14 个数据包级别增量操作和时间扰动操作的动作空间。代理以黑盒检测模型的二进制判别结果为奖励，根据强化学习策略从动作空间中选择下一个修改动作，迭代修改原始流量样本，直到成功欺骗目标模型或超出最大访问限制。

Cheng 等[43]从基于策略梯度的序列生成模型 SeqGAN[55]中获得灵感，提出了 Attack-GAN 攻击。作者将对抗流量的生成建模为序列决策过程，生成器相当于强化学习中的代理，生成的字节为当前状态，动作空间为全部可能字节。生成器将以判别器的梯度信息作为奖励，利用蒙特卡罗树搜索算法来选定下一字节，并通过仅修改不影响恶意功能的字节和遵守网络协议头部字段值域范围来满足保持恶意功能和遵守网络协议规范的约束。

文献[44]提出了一种新的端到端基于时间的对抗流量重构攻击——TANTRA。TANTRA 利用长短时记忆网络学习良性数据包的时间差分布特点，从而实现在不改变恶意数据包内容的前提下，通过改变恶意流量数据包的时间差分布绕过检测。在 CIC-IDS2017[56]上的实验实现了 99.99%的平均成功率。

可以发现，在 NIDS 领域的对抗攻击方法从最开始简单的迁移计算机视觉领域的工作，逐步发展为结合 NIDS 具体应用场景开发新型对抗攻击方法。虽然很多研究工作都声称自己可以实现黑盒场景下的 NIDS 对抗攻击，但或者因为所使用的是 NSL-KDD 的特征数据集而依赖目标模型的特征

集，或是需要了解目标模型分类得分都仅能实现灰盒攻击，但并不代表这些工作完全不可采纳。由于恶意流量的恶意特征往往比较明显，且不同检测模型使用的特征空间通常具有一定的重复性，而攻击者作为经验丰富的恶意专家很可能拥有关于检测常用恶意特征的先验知识。因此攻击者完全可能在不了解目标模型确切特征空间的条件下，基于领域先验知识选择典型恶意流量特征施加扰动，进而实现对黑盒 NIDS 模型的对抗攻击。

从输出结果角度分析，流量空间攻击则显得更有价值。流量空间攻击的重要环节是生成对抗性恶意流量，主要包括以下几种方法：1) 直接在原始流量上施加数据包增量和时延操作，并利用代理实现扰动，如文献[31-32,41-42,44]；2)利用其他组件从对抗特征构造对抗流量，同时仅修改非功能性特征，而不损害原始流量功能，如文献[39]；3)将对抗特征传递给控制端，使其根据需求重新生成恶意流量，这需要对源码施加复杂的修改，如文献[40, 54]。

6　对抗防御

随着对抗攻击方法研究的不断深入，为增强 ML 模型的稳健性，抵御对抗攻击的安全威胁，研究者提出了多种防御方法，根据修正对象的不同可分为针对模型的修正、针对数据的修正和使用附加网络 3 种类型。相关工作整理如表 2 所示。

1) 针对模型的修正

防御蒸馏。在防御蒸馏模型[57]中，选择 2 个相同的模型作为教师和学生模型，将原始分类模型学到的信息迁移到小型网络模型中，从而实现了梯度遮掩。防御蒸馏可以有效抵抗一些基于梯度的小幅度扰动的对抗攻击，但在未知模型函数或黑盒攻击的情况下，该防御方法失效。

深度压缩网络（DCN, deep contractive networks）。文献[58]提出了一种融合了平滑惩罚的端到端训练模型 DCN，从而在保证不会显著降低性能的前提下，增加了网络对对抗样本的稳健性。与这种去噪的思想类似，文献[59]利用去噪自编码器构建 NIDS，并在数据输入模型前应用多个随机掩码增加输入数据的噪声扰动，实现了相当于 Kitsune-GMM79 倍的检测率，在对抗环境中的检测率也是 Kitsune-GMM 的 3.73 倍。

集成分类模型。文献[60]提出一种层次集成的 NIDS。其中每个弱分类器都使用不同特征集，且后置分类器仅处理前置分类器识别为良性的样本，从而保证被识别为良性的样本可经过全部弱分类器。实验证明该防御方法可 100%抵抗基于最邻近算法的对抗攻击[62]。但这种方法会显著增加防御成本。

基于邻域分类的防御。Cao 等[63]基于对抗样本的分布接近于分类边界这一观察，提出以基于区域分类的防御方法。具体来说，对于每个待预测样本，在以其为中心的超立方体邻域范围内随机选取若

表 2　对抗防御方法

文献	年份	类别	方法	评估指标	数据集	可以抵抗的攻击
文献[57]	2016	模型修正	防御蒸馏	ASR 变化，扰动特征数量	MNIST, CIFAR-10	文献[18]
文献[58]	2015	模型修正	DCN	测试误差，平均扰动大小	MNIST	文献[4]
文献[59]	2020	模型修正	DAE+随机掩码	TPR, FPR	CIC-IDS2017	文献[41]
文献[60]	2020	模型修正	层次集成分类	混淆矩阵	网络扫描流量数据集[61]	文献[62]
文献[63]	2017	模型修正	邻域分类	ASR 变化,平均扰动大小	MNIST, CIFAR-10	文献[18-21],BIM
文献[64-65]	2020	数据修正	对抗训练（文献[21,24,49], BIM, PGD），特征选择	accuracy, precision, TPR, AUC	UNSW-NB 15	文献[19-21],BIM, PGD
文献[66]	2021	数据修正	对抗训练（文献[18-19]）	F1-score	ICS 流量数据集[67]	文献[18]
文献[68]	2020	数据修正	非稳健特征删减	ASR 变化,平均扰动大小	Kitsunet dataset	文献[38]
文献[69]	2018	附加网络	Defence-GAN	accuracy, ROC	MNIST, F-MNIST	文献[18,21]
文献[70]	2019	附加网络	APE-GAN	对抗样本的分类错误率	MNIST,CIFAR-10,ImageNet	文献[4,18-21]
文献[71]	2017	附加网络	MagNet	accuracy	MNIST, CIFAR-10	文献[18-21], BIM
文献[72]	2017	附加网络	基于分类子网的检测	accuracy	CIFAR-10	文献[18,20], BIM
文献[73]	2020	附加网络	基于分类子网的检测	precision, recall, F1-score	CIC-IDS-2017	文献[18,21], BIM, PGD

干个样本，采用多数表决方式选择预测标签最多的作为待预测样本最终的标签。通过在 MNIST 和 CIFAR-10 数据集上的实验证明，该方法可以防御 FGSM、C&W、JSMA、BIM 和 DeepFool 等多种先进的攻击手法，同时不牺牲分类精度。

2) 针对数据的修正

对抗训练。对抗训练通过在模型训练数据集添加预先构造的对抗样本，提升模型针对对抗样本的稳健性。根据加入对抗样本的不同，又可进一步分为 FGSM 对抗训练[23]、PGD 对抗训练[74]和集成对抗训练[75]防御方法。前两者是仅利用某种特定攻击方法快速构造大量对抗样本进行对抗训练，而集成对抗训练则是利用多种类型的对抗样本来对原始数据集进行数据增强。对抗训练泛化性能较差，仅能防御已知的攻击类型，因此能够发挥的作用有限。

在 NIDS 领域，文献[64-66]都使用了对抗训练的方法来加强 IDS 模型对于对抗攻击的稳健性，并在多种数据集、多种攻击方法、多种分类器模型下进行了全面细致的评估。文献[64]还证明基于主成分分析的特征约减也能够显著提高 IDS 的稳健性。

非稳健特征约减。文献[76]提出了一种特征压缩的方法来检测对抗样本，其基本思想是巨大的特征输入空间为攻击者构建对抗样本提供了很大的空间，因此防御者可以通过“压缩”不必要的输入特征来减少攻击者的自由度，限制其攻击行为。文献[68]借鉴了这一思想，通过删除一些稳健性得分较低的特征来防御潜在的对抗攻击。通过删除 20%分数较低的特征，与对抗训练和简单的特征选择相比，该方法能够实现更好的防御性能，同时分类器的检测性能不会受到显著影响（ΔF1-score<5%）。

3) 使用附加网络检测

基于 GAN 的防御。文献[69]阐述了一种基于 GAN 框架的防御——Defence-GAN 训练生成器学习原始样本分布，从而在测试阶段搜索接近于对抗样本的原始图像将对抗样本转化为正常样本，降低对抗扰动噪声带来的影响。Defence-GAN 可防御多种对抗攻击，具有一定的泛化能力。APE-GAN[70]则为生成器设计了一个混合损失函数，使其以对抗样本为输入，学习生成与原始图像相似且消除了对抗性扰动的重建图像，从而缓解对抗样本影响。

基于流形的防御。MagNet[71]基于流形假设，通过拒绝或重塑对抗样本来保护目标模型。MagNet 由检测模块和重塑模块组成。检测模块测量测试样本与正常流形之间的距离，基于重构误差和概率分歧检测对抗样本。重塑模块使用自动编码器将对抗样本推向正常流形，使之成为合法样本。实验证明 MagNet 对于黑盒/灰盒攻击有较好的防御效果，且不局限于某种对抗攻击，具有相当强的泛化能力。

基于分类子网的防御。Metzen 等[72]通过为目标神经网络分类器的某一层分支添加检测器子网络来识别对抗样本。检测器以分类器的中间表示为输入，利用类似对抗训练的方式动态获得对抗样本以学习原始样本和对抗样本的差异。Pawlicki 等[73]将类似的思想迁移到基于神经网络的 NIDS 模型上，作者不再局限于单层神经元激活，而是以 NIDS 神经网络模型的全部神经元输出为样本，训练检测器识别对抗样本。实验证明，该方法可实现 70%～99%的对抗样本检测召回率。

7　讨论与展望

机器学习技术的迅速发展使其在 NIDS 领域得到广泛应用的同时也暴露了一定的安全隐患。针对其对对抗样本的脆弱性问题，安全研究人员展开了一系列攻击防御方法研究。从简单的方法移植到综合考虑 NIDS 领域特有约束，攻击和防御方法在相互博弈中不断发展。通过总结分析，本文提出了 3 点未来研究方向。

1) 流量空间攻击研究

当前针对流量空间攻击方法的研究仍然相对较少，且在 NIDS 领域，只有能够生成对抗流量样本的攻击才能产生实际安全威胁。由于从特征向量到流量样本的映射十分复杂，因此现有的流量空间攻击往往采用直接在原始流量样本上施加扰动的方法来规避逆映射困境。相信就像在计算机视觉领域的对抗攻击是从特征空间攻击逐步发展到问题空间攻击的一样，其在 NIDS 领域的发展也将逐渐更加适应真实攻击场景。

2) 对抗防御方法研究

目前针对 NIDS 领域的对抗攻击防御研究主要围绕对抗训练和选择更复杂稳健的模型架构展开。但两者都需要较大的额外开销，在 NIDS 领域的适用性还有待评估。如何结合恶意流量对抗样本的分布和数据特征，设计开销小、速度快、准确性高且具有一定适应性的对抗样本检测或防御方法也是

值得进一步研究的课题。

3) 标准流量数据集构建

一个新颖、全面、真实的数据集是在 NIDS 领域开展研究的重要保障。当前研究广泛使用的 NSL-KDD、CIC-IDS、CTU-13 等数据集往往无法覆盖日新月异的新型攻击手法和逃避技术，无法有效模拟真实的场景。同时关于构建高对抗性数据集的研究[77-78]也比较罕见，这也进一步导致了当前防御方法的评估莫衷一是的局面。因此如何构建一个标准、良好且具有高对抗性的数据集是一个亟待解决的课题。

8 结束语

在 NIDS 这一具有高对抗性的攻防领域，对抗攻击的存在将严重威胁网络和用户安全。本文从攻防 2 个视角全面调研了 NIDS 领域的对抗攻击和防御方法。首先，本文提出了在 NIDS 领域应用对抗攻击特有的约束和挑战。然后，构建多维分类法，结合 IDS 场景需求从攻防 2 个角度整理对比现有研究成果。最后，总结当前研究现状，并探讨 NIDS 领域对抗攻击的未来发展方向。

参考文献：

[1] ANDERSON J P. Computer security threat monitoring and surveillance[J]. Technical Report James P Anderson Co Fort Washington Pa, 1980: 56.

[2] SINCLAIR C, PIERCE L, MATZNER S. An application of machine learning to network intrusion detection[C]//Proceedings of the 15th Annual Computer Security Applications Conference (ACSAC'99). Piscataway: IEEE Press, 1999: 371-377.

[3] WU S X, BANZHAF W. The use of computational intelligence in intrusion detection systems: a review[J]. Applied Soft Computing, 2010, 10(1): 1-35.

[4] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[C]//International Conference on Learning Representations. [S.l.:s.n.], 2014.

[5] PAPERNOT N, MCDANIEL P, GOODFELLOW I. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples[J]. arXiv Preprint, arXiv:1605.07277, 2016.

[6] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial examples in the physical world[C]//International Conference on Learning Representations. [S.l.:s.n.], 2017.

[7] ALZANTOT M, SHARMA Y, ELGOHARY A, et al. Generating natural language adversarial examples[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018: 2890-2896.

[8] QIN Y, CARLINI N, GOODFELLOWI, et al. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition[C]//Proceedings of the 36th International Conference on Machine Learning. Australia: PMLR, 2019: 5231-5240.

[9] FREDRIKSON M, LANTZ E, JHA S, et al. Privacy in pharmacogenetics: an end-to-end case study of personalized warfarin dosing[C]// Proceedings of the USENIX Security Symposium. Berkeley: USENIX Association, 2014: 17-32.

[10] PAPERNOT N, MCDANIEL P, GOODFELLOW I, et al. Practical black-box attacks against machine learning[J].arXiv Preprint, arXiv:1602.02697, 2016.

[11] SHARIF M, BHAGAVATULA S, BAUER L, et al. Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition[C]//Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2016: 1528-1540.

[12] 张玉清, 董颖, 柳彩云, 等. 深度学习应用于网络空间安全的现状、趋势与展望[J]. 计算机研究与发展, 2018, 55(6): 1117-1142.
ZHANG Y Q, DONG Y, LIU C Y, et al. Situation, trends and prospects of deep learning applied to cyberspace security[J]. Journal of Computer Research and Development, 2018, 55(6): 1117-1142.

[13] AKHTAR N, MIAN A. Threat of adversarial attacks on deep learning in computer vision: a survey[J]. IEEE Access, 2018, 6: 14410-14430.

[14] MARTINS N, CRUZ J M, CRUZ T, et al. Adversarial machine learning applied to intrusion and malware scenarios: a systematic review[J]. IEEE Access, 2020, 8: 35403-35419.

[15] ROSENBERG I, SHABTAI A, ELOVICI Y, et al. Adversarial learning in the cyber security domain[J]. arXiv Preprint, arXiv: 2007.02407, 2020.

[16] 段广晗, 马春光, 宋蕾,等. 深度学习中对抗样本的构造及防御研究[J]. 网络与信息安全学报, 2020, 6(2): 1-11.
DUAN G H, MA C G, SONG L, et al. Research on structure and defense of adversarial example in deep learning[J]. Chinese Journal of Network and Information Security, 2020, 6(2): 1-11.

[17] YUAN X Y, HE P, ZHU Q L, et al. Adversarial examples: attacks and defenses for deep learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(9): 2805-2824.

[18] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]//International Conference on Learning Representations. [S.l.:s.n.], 2015.

[19] PAPERNOT N, MCDANIEL P, JHA S, et al. The limitations of deep learning in adversarial settings[C]//Proceedings of 2016 IEEE European Symposium on Security and Privacy (EuroS&P). Piscataway: IEEE Press, 2016: 372-387.

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

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

[22] MOOSAVI-DEZFOOLI S M, FAWZI A, FAWZI O, et al. Universal adversarial perturbations[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 86-94.

[23] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial machine learning at scale[C]//International Conference on Learning Representation. [S.l.:s.n.], 2017.

[24] PIERAZZI F, PENDLEBURY F, CORTELLAZZI J, et al. Intriguing properties of adversarial ML attacks in the problem space[C]//Proceedings of 2020 IEEE Symposium on Security and Privacy (SP).

Piscataway: IEEE Press, 2020: 1332-1349.

[25] SCHOLKOPF B, MIKA S, BURGES C J C, et al. Input space versus feature space in kernel-based methods[J]. IEEE Transactions on Neural Networks, 1999, 10(5): 1000-1017.

[26] RIGAKI M. Adversarial deep learning against intrusion detection classifiers[EB]. 2017.

[27] WANG Z. Deep learning-based intrusion detection with adversaries[J]. IEEE Access, 2018, 6: 38367-38384.

[28] IBITOYE O, SHAFIQ O, MATRAWY A. Analyzing adversarial attacks against deep learning for intrusion detection in IoT networks[C]//Proceedings of 2019 IEEE Global Communications Conference (GLOBECOM). Piscataway: IEEE Press, 2019: 1-6.

[29] KLAMBAUER G, UNTERTHINER T, MAYR A, et al. Self-normalizing neural networks[C]//Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017). New York: Curran Associates Inc, 2017: 1-8.

[30] NOVO C, MORLA R. Flow-based detection and proxy-based evasion of encrypted malware c2 traffic[C]//Proceedings of the 13th ACM Workshop on Artificial Intelligence and Security. New York: ACM Press, 2020: 83.

[31] CHERNIKOVA A, OPREA A. Fence: feasible evasion attacks on neural networks in constrained environments[J]. arXiv Preprint, arXiv: 1909.10480, 2019.

[32] 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.

[33] YANG K C, LIU J Q, ZHANG C, et al. Adversarial examples against the deep learning based network intrusion detection systems[C]//Proceedings of MILCOM 2018 - 2018 IEEE Military Communications Conference (MILCOM). Piscataway: IEEE Press, 2018: 559-564.

[34] LIN Z, SHI Y, XUE Z. IDSGAN: Generative adversarial networks for attack generation against intrusion detection[J]. arXiv Preprint, arXiv:1809.02077, 2018.

[35] SHU D L, LESLIE N O, KAMHOUA C A, et al. Generative adversarial attacks against intrusion detection systems using active learning[C]//Proceedings of the 2nd ACM Workshop on Wireless Security and Machine Learning. New York: ACM Press, 2020:1-6.

[36] AIKEN J, SCOTT-HAYWARD S. Investigating adversarial attacks against network intrusion detection systems in SDNs[C]//Proceedings of 2019 IEEE Conference on Network Function Virtualization and Software Defined Networks(NFV-SDN). Piscataway: IEEE Press, 2019: 1-7.

[37] SHEFFEY S, ADERHOLDT F. Improving meek with adversarial techniques[C]//Proceedings of the 9th USENIX Workshop on Free and Open Communications on the Internet. Santa Clara: USENIX Association, 2019: 1-10.

[38] ALHAJJAR E, MAXWELL P, BASTIANN. Adversarial machine learning in network intrusion detection systems[J]. Expert Systems With Applications, 2021, 186: 115782.

[39] KUPPA A, GRZONKOWSKI S, ASGHAR M R, et al. Black box attacks on deep anomaly detectors[C]//Proceedings of the 14th International Conference on Availability, Reliability and Security. New York: ACM Press, 2019: 1-10.

[40] RIGAKI M, GARCIA S. Bringing a GAN to a knife-fight: adapting malware communication to avoid detection[C]//Proceedings of 2018 IEEE Security and Privacy Workshops (SPW). Piscataway: IEEE Press, 2018: 70-75.

[41] HASHEMI M J, CUSACK G, KELLER E. Towards evaluation of NIDSs in adversarial setting[C]//Proceedings of the 3rd ACM CoNEXT Workshop on BigDAta, Machine Learning and Artificial Intelligence for Data Communication Networks. New York: ACM Press, 2019: 14-21.

[42] WU D, FANG B X, WANG J N, et al. Evading machine learning botnet detection models via deep reinforcement learning[C]//Proceedings of ICC 2019 - 2019 IEEE International Conference on Communications(ICC). Piscataway: IEEE Press, 2019: 1-6.

[43] CHENG Q, ZHOU S, SHEN Y, et al. Packet-level adversarial network traffic crafting using sequence generative adversarial networks[J]. arXiv Preprint, arXiv: 2103.04794, 2021.

[44] SHARON Y, BEREND D, LIU Y, et al. Tantra: timing-based adversarial network traffic reshaping attack[J]. arXiv Preprint, arXiv: 2103.06297, 2021.

[45] 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.

[46] CLEMENTS J, YANG Y, SHARMA A, et al. Rallying adversarial techniques against deep learning for network security[J]. arXiv Preprint, arXiv:1903.11688, 2019.

[47] MIRSKY Y, DOITSHMAN T, ELOVICI Y, et al. Kitsune: an ensemble of Autoencoders for online network intrusion detection[C]//Proceedings of 2018 Network and Distributed System Security Symposium. Reston: Internet Society, 2018: 18-21.

[48] GARCÍA S, GRILL M, STIBOREK J, et al. An empirical comparison of botnet detection methods[J]. Computers & Security, 2014, 45: 100-123.

[49] 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. New York：ACM Press, 2017: 15-26.

[50] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein generative adversarial networks[C]//Proceedings of the 34th International Conference on Machine Learning. Australia: PMLR, 2017(70): 214-223.

[51] CHOI Y, CHOI M, KIM M, et al. StarGAN: unified generative adversarial networks for multi-domain image-to-image translation[C]// Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 8789-8797.

[52] MOUSTAFA N, SLAY J. UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)[C]//Proceedings of 2015 Military Communications and Information Systems Conference(MilCIS). Piscataway: IEEE Press, 2015: 1-6.

[53] LI D, MUKHOPADHYAY M, DUNSON D B. Efficient manifold and sub-space approximations with spherelets[J]. arXiv Preprint, arXiv: 1706.08263, 2017.

[54] LI J, ZHOU L, LI H X, et al. Dynamic traffic feature camouflaging via generative adversarial networks[C]//Proceedings of 2019 IEEE Conference on Communications and Network Security(CNS). Piscataway: IEEE Press, 2019: 268-276.

[55] YU L, ZHANG W, WANG J, et al. SeqGAN: sequence generative adversarial nets with policy gradient[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2017: 2852-2858.

[56] SHARAFALDIN I, HABIBI L A, GHORBANI A A. Toward generat-

ing a new intrusion detection dataset and intrusion traffic characterization[C]//Proceedings of the 4th International Conference on Information Systems Security and Privacy. [S.l.]: SciTeOress, 2018: 108-116.

[57] PAPERNOT N, MCDANIEL P, WU X, et al. Distillation as a defense to adversarial perturbations against deep neural networks[C]// Proceedings of 2016 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2016: 582-597.

[58] GU S, RIGAZIO L. Towards deep neural network architectures robust to adversarial examples[C]//International Conference on Learning Representations. [S.l.:s.n.], 2015.

[59] HASHEMI M J, KELLER E. Enhancing robustness against adversarial examples in network intrusion detection systems[C]//Proceedings of 2020 IEEE Conference on Network Function Virtualization and Software Defined Networks(NFV-SDN). Piscataway: IEEE Press, 2020: 37-43.

[60] DE L M J, COTTON C. A network security classifier defense: against adversarial machine learning attacks[C]//Proceedings of the 2nd ACM Workshop on Wireless Security and Machine Learning. New York: ACM Press, 2020: 67-73.

[61] VENKATESAN S, SUGRIM S, IZMAILOV R, et al. On detecting manifestation of adversary characteristics[C]//Proceedings of MILCOM 2018 - 2018 IEEE Military Communications Conference (MILCOM). Piscataway: IEEE Press, 2018: 431-437.

[62] DE L M J, COTTON C. Adversarial machine learning for cybersecurity[J]. JISAR, 2019, 12(1): 26.

[63] CAO X Y, GONG N Z. Mitigating evasion attacks to deep neural networks via region-based classification[C]//Proceedings of the 33rd Annual Computer Security Applications Conference. New York: ACM Press, 2017: 278-287.

[64] KHAMIS R A, SHAFIQ M O, MATRAWY A. Investigating resistance of deep Learning-based ids against adversaries using Min-max optimization[C]//Proceedings of ICC 2020 - 2020 IEEE International Conference on Communications (ICC). Piscataway: IEEE Press, 2020: 1-7.

[65] KHAMIS R A, MATRAWY A. Evaluation of adversarial training on different types of neural networks in deep learning-based IDSs[C]//Proceedings of 2020 International Symposium on Networks, Computers and Communications (ISNCC). Piscataway: IEEE Press, 2020: 1-6.

[66] ANTHI E, WILLIAMS L, RHODE M, et al. Adversarial attacks on machine learning cybersecurity defences in Industrial Control Systems[J]. Journal of Information Security and Applications, 2021, 58: 102717.

[67] PAN S Y, MORRIS T, ADHIKARI U. Classification of disturbances and cyber-attacks in power systems using heterogeneous time-synchronized data[J]. IEEE Transactions on Industrial Informatics, 2015, 11(3): 650-662.

[68] HAN D, WANG Z, ZHONG Y, et al. Practical traffic-space adversarial attacks on learning-based NIDSs[J]. arXiv Preprint, arXiv: 2005.07519, 2020.

[69] SAMANGOUEI P, KABKAB M, CHELLAPPA R. Defense-gan: protecting classifiers against adversarial attacks using generative models[C]//International Conference on Learning Representations. [S.l.:s.n.], 2018.

[70] JIN G Q, SHEN S W, ZHANG D M, et al. APE-GAN: adversarial perturbation elimination with GAN[C]//Proceedings of ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2019: 3842-3846.

[71] MENG D Y, CHEN H. MagNet: a two-pronged defense against adversarial examples[C]//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2017: 135-147.

[72] METZEN J H, GENEWEIN T, FISCHER V, et al. On detecting adversarial perturbations[C]//Proceedings of Internet Conference on Learning Representations. [S.l.:s.n.], 2017.

[73] PAWLICKI M, CHORAŚ M, KOZIK R. Defending network intrusion detection systems against adversarial evasion attacks[J]. Future Generation Computer Systems, 2020, 110: 148-154

[74] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv Preprint, arXiv: 1706.06083, 2017.

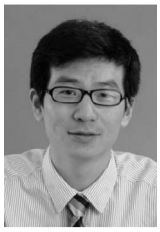
[75] TRAM`ER F, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial training: attacks and defenses[J]. arXiv Preprint, arXiv: 1705.07204, 2017.

[76] XU W L, EVANS D, QI Y J. Feature squeezing: detecting adversarial examples in deep neural networks[C]//Proceedings of 2018 Network and Distributed System Security Symposium. Reston: Internet Society, 2018: 18-21.

[77] VENTURI A, APRUZZESE G, ANDREOLINI M, et al. DReLAB - Deep reinforcement learning adversarial botnet: a benchmark dataset for adversarial attacks against botnet Intrusion Detection Systems[J]. Data in Brief, 2021, 34: 106631.

[78] HOMOLIAK I, MALINKA K, HANACEK P. ASNM datasets: a collection of network attacks for testing of adversarial classifiers and intrusion detectors[J]. IEEE Access, 2020, 8: 112427-112453.

[作者简介]



刘奇旭（1984− ），男，江苏徐州人，博士，中国科学院信息工程研究所研究员，中国科学院大学教授，主要研究方向为网络攻防技术、网络安全评测。

王君楠（1995− ），女，吉林省吉林市人，中国科学院大学博士生，主要研究方向为机器学习、机器学习安全和恶意流量检测。

尹捷（1991− ），女，重庆人，博士，中国科学院信息工程研究所工程师，主要研究方向为网络攻防技术、恶意代码分析。

陈艳辉（1996− ），男，山东潍坊人，中国科学院大学博士生，主要研究方向为网络攻防技术和恶意软件分析与检测。

刘嘉熹（1997− ），女，山东淄博人，中国科学院大学博士生，主要研究方向为恶意代码分析。