doi :10.3969/j.issn.1671-1122.2019.02.001

基于 GAN 的网络攻击检测研究综述

傅建明[1,2]，黎琳[1]，郑锐[1]，苏日古嘎[1]

(1. 武汉大学国家网络安全学院，湖北武汉 430072；
2. 武汉大学空天信息安全与可信计算教育部重点实验室，湖北武汉 430072)

摘　要：生成式对抗网络（Generative Adversarial Network，GAN）是近年来深度学习领域的一个重大突破，是一个由生成器和判别器共同构成的动态博弈模型。其“生成”和“对抗”的思想获得了广大科研工作者的青睐，满足了多个研究领域的应用需求。受该思想的启发，研究者们将 GAN 应用到网络安全领域，用于检测网络攻击，帮助构建智能有效的网络安全防护机制。文章介绍了 GAN 的基本原理、基础结构、理论发展和应用现状，着重从网络攻击样本生成、网络攻击行为检测两大方面研究了其在网络攻击检测领域的应用现状。

关键词：GAN；生成器；判别器；网络攻击；网络安全

中图分类号：TP309　**文献标识码：**A　**文章编号：**1671-1122（2019）02-0001-09

中文引用格式：傅建明，黎琳，郑锐，等 . 基于 GAN 的网络攻击检测研究综述 [J]. 信息网络安全，2019,19（2）:1–9.

英文引用格式：FU Jianming, LI Lin, ZHENG Rui, et al. Survey of Network Attack Detection Based on GAN[J]. Netinfo Security, 2019,19（2）: 1–9.

Survey of Network Attack Detection Based on GAN

FU Jianming[1, 2], LI Lin[1], ZHENG Rui[1], Suriguga[1]

(1. *School of Cyber Science and Engineering, Wuhan University, Wuhan Hubei* 430072, *China*; 2. *Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, Wuhan University, Wuhan Hubei* 430072, *China*)

Abstract: Generative adversarial network (GAN) is a major breakthrough in the field of deep learning in recent years. It is a dynamic game model composed of generator and discriminator. Its ideas of “generation” and “confrontation” have won the favor of the vast number of scientific researchers and met the application needs of many research fields. Inspired by the ideas, researchers applied GAN to the field of network security to detect network attacks and help build an intelligent and effective network security protection mechanism. This paper introduces the basic principle, infrastructure, theoretical development and application status of GAN, and focuses on the application status of GAN in the field of network attack detection from two aspects of network attack sample generation and network attack behavior detection.

Key words: GAN; generator; discriminator; network attack; network security

收稿日期：2018-9-28

基金项目：国家自然科学基金 [61373168]；国家自然科学基金联合基金 [U1636107]

作者简介：傅建明（1969—），男，湖南，教授，博士，主要研究方向为系统安全、移动安全；黎琳（1996—），女，贵州，硕士研究生，主要研究方向为网络安全；郑锐（1992—），男，河南，博士研究生，主要研究方向为网络安全；苏日古嘎（1993—），女，内蒙古，硕士研究生，主要研究方向为网络安全。

通信作者：傅建明 3578428633@qq.com

0 引言

近年来人工智能蓬勃发展，深度学习作为其最为活跃的分支之一，在文本、语音、计算机视觉等领域取得了许多重要进展。其中，生成式对抗网络（Generative Adversarial Network，GAN）作为一种深度生成式模型，成为众多科研工作者的研究热点。GAN 本质上是一个由生成器（Generator）和判别器（Discriminator）构成的动态博弈模型。由于其强大的生成能力和新颖的对抗思想，GAN 能够满足计算机图像与视觉、语音处理、语言处理、网络安全等多个研究领域的应用需求，帮助解决图像生成、图像融合 [1]、机器翻译、语音生成 [2]、文本分类 [3] 等问题。

本文第 1 章介绍了GAN 的基本原理、基础结构及其理论发展和应用现状；第 2 章介绍了GAN 在网络攻击检测领域的应用现状，包括网络攻击样本生成和网络攻击行为检测两个方面；第 3 章对全文进行总结。

1 GAN 概述

1.1 基本原理和基础结构

GAN 最早由 GOODFELLOW[4] 等人于 2014 年提出，其本质上是生成器和判别器之间的动态最小最大博弈游戏。生成器捕获真实数据分布，根据给定的噪声生成新的对抗样本；判别器评估所接收的输入是真实数据而不是生成器所生成的对抗样本的概率。如果概率大于 0.5，则为真实数据；小于 0.5，则为对抗样本。在这种动态博弈训练过程中，生成器的目的是增大判别器犯错的概率，判别器的目的是将真实数据和对抗样本分开。两者不断训练以提高自身的生成能力和判别能力，直至判别器不足以区分真实数据和对抗样本，即生成器能够生成最为逼真的对抗样本，判别器无论对于真实样本还是对抗样本，输出的概率都为 0.5，生成器和判别器之间达到了一个纳什平衡。

GAN 的基本结构和训练过程如图 1 所示。图 1 中，x 表示真实数据，z 表示输入生成器的随机噪声，$G(z)$ 是生成器所生成的对抗样本。如果判别器的输入是 x，则判别为真并标注为 1；如果输入是 $G(z)$，则判别为假并标注为 0。在迭代对抗训练过程中，生成器希望自己生成的样本无限接近真实数据的数据分布，也就是说生成器希望 $D(G(z))$（判别器判别 $G(z)$ 为真的概率）尽可能地大。当达到理想状态时，$G(z)$ 和 x 在数据分布上不具有差异性，判别器难以正确判断其输入。

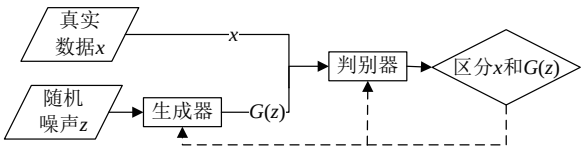

图 1 GAN 基本结构和训练过程

1.2 GAN 的理论发展

GAN 自面世以来就获得了极大关注，研究者们的不断探索使得 GAN 在原有理论的基础上不断拓展更新。

1.2.1 Improved GAN

训练 GAN 的目的是要达到两方博弈的纳什均衡，但是仅仅采用随机梯度下降的方法只能降低损失函数的值而无法达到纳什均衡。对此，Improved GAN[5] 引入特征匹配（Feature Matching）和批次判别（Minibatch Discrimination）两种方法来提高模型的收敛度。判别器从输入层到输出层逐层卷积和池化，图片信息也逐渐损失，因此中间层比输出层能够得到更好更多的样本分布信息。特征匹配的目的是利用损失函数，使得 GAN 生成的样本经过判别器中间层得到的特征图和真实样本经过判别器中间层得到的特征图尽可能相同。因此，Improved GAN 中的生成器生成的对抗样本更加符合真实样本的分布规律。但 Improved GAN 依然不能保证一定能够达到纳什均衡。Improved GAN 为了避免生成器过快停止训练更新，一旦生成一个被判别器认定为“真”的样本，就会一直生成同样的内容，从而导致因缺乏样本多样性而产生的所谓模式坍塌（Mode Collapse）问题。对此 Improved GAN 引入批次判别方法，让判别器每次都同时判别一批样本，使得生成器能够生成多样性更强的对抗样本，从而有效解决模式坍塌问题。

1.2.2 EBGAN

ZHAO[6] 等人丢弃了原先的概率视角，选择从能量视角构建一个基于能量的生成式对抗网络（Energy-based GAN，EBGAN）。该网络中判别器被看成是一个能量函数，该函数赋予靠近真实数据的区域低能量而赋予其他非真实数据的区域高能量。生成器的目的是为了生成对抗样本，判别器的目的是赋予这些对抗样本高能量。将判别器看作能量函数的好处在于能够融合使用更多不同的网络结构（如 CNN、RNN、DNN）和损失函数。相较于只有逻辑输出的二元分类器而言，EBGAN 在训练过程中更为灵活多变。在 MNIST、LSUN 以及 CELEBA 数据集上的实验证明，EBGAN 在训练稳定性和样本质量方面都优于传统 GAN，且生成器和判别器所采用的基础结构和参数对于自身网络的影响小于传统 GAN。然而，由于 EBGAN 选取的损失函数和传统 GAN 类似，在训练过程中会出现传统 GAN 所面临的问题（如模式坍塌），因而也无法将判别器训练到最优状态。

1.2.3 Wasserstein GAN

针对传统 GAN 训练的不稳定性和脆弱性问题，ARJOVSKY[7] 等人提出了 Wasserstein GAN（WGAN）。与传统 GAN 相比，WGAN 大大降低了 GAN 训练过程中模式坍塌出现的概率。同时，WGAN 算法不需要维持生成器和判别器在训练过程中步长的绝对平衡，也不需要对基础网络结构进行细致的设计，提高了训练过程的稳定性以及对抗样本的质量。ARJOVSKY[7] 等人在 LSUN-Bedrooms 数据集上设计了图片生成实验来验证 WGAN 相对于传统 GAN 的优越性，实验观察到当生成器选择 CNN（卷积神经网络）时，WGAN 具备更强的健壮性。然而，尽管 WGAN 在 GAN 训练的稳定性上前进了一大步，但依然存在生成低质量样本或在某些集合上难以收敛的问题。GULRAJANI[8] 等人发现上述问题的出现是由于 WGAN 中使用了权重剪裁，使得判别器必须处于 1-Lipschitz 空间 [7] 中。因此，GULRAJANI[8] 等人采用梯度惩罚来替代权重裁剪，使得经过修改的 WGAN 与原来的 WGAN 算法相比，能够加快收敛速度，提高对抗样本质量，且无需调整超参数便能训练不同网络结构的生成器。修改过的 WGAN 在 CIFAR-10 和 LSUN-Bedrooms 数据集上均获得了高质量的对抗样本。

1.3 GAN 的应用领域

GAN 自 2014 年出现以来，经过众多科研人员的研究与探索，不仅在理论上获得了长足的进步，也被用于处理文本、图片、音乐、视频、音频等数据，从而解决了一些重要问题，如图像生成、对话生成、文本分类、文本生成图像等。本节从图像视觉、语音、文本等领域来阐述 GAN 的应用。

1.3.1 图像视觉领域

在图像生成方面，LEDIG[9] 等人提出了 SRGAN 模型，该模型能够将低分辨率的图像放大 4 倍变为高分辨率的图像，同时保持图像的高保真度和纹理细节。为了达到上述目的，LEDIG[9] 等人提出了由对抗性损失和内容损失构成的感知损失函数，同时用参数化的深度残差网络作为生成器，用 VCG 网络作为判别器。采用 MOS（Mean Opinion Score，平均意见分）的实验证明，相比通过现有图像超分辨率方法所获得的图像，SRGAN 能够得到更为接近原始图像细节和保真度的高分辨率图像。

在图像融合方面，WU[10] 等人结合泊松 - 高斯方程和 GAN 提出了 GP-GAN 算法。GAN 虽然常用于生成自然图片，但难以捕获图片细节如纹理和边界。基于梯度的高斯算法能够生成高分辨率的图像，但其人工制造的痕迹过于严重。GP-GAN 将 GAN 和基于梯度的高斯算法融合之后取长补短，克服了两个方法的缺点，从而可以实现高分辨率的图像融合。

在图像修补方面，PATHAK[11] 等人提出无监督视觉特征算法。该算法采用 GAN 的思想，将编码器和解码器分别作为判别器和生成器，提出了一个基于上下文像素预测的自编码器。该自编码器基于图像缺失部分的周围环境，利用卷积神经网络生成图像缺失区

域的内容。该算法在分类、物体检测以及分段任务等方面的应用也是十分有效的。

1.3.2 语音领域

在语音领域，C-RNN-GAN[12] 可以说是最早用以生成音乐的 GAN 模型。该模型结合了 RNN 和 GAN，将随机噪声作为输入以生成多样的旋律，并从复调、尺度一致性、重复性、音调跨度等方面来评估生成的旋律。但是该模型缺乏条件机制，无法根据给定的前奏和和弦序列生成音乐。SeqGAN 模型 [13] 首先基于策略梯度训练生成器，然后通过蒙特卡洛搜索得到策略梯度的反馈奖励信号，再基于反馈信号训练生成器提升生成能力。实验表明，SeqGAN 模型无论是在语音生成方面还是在音乐生成方面都优于传统的 RNN 模型。

1.3.3 文本领域

在文本领域，LI[14] 等人提出可以用 GAN 学习并表示对话之间的隐式关系，根据这些关系重构对话文本。此外，ZHANG[15] 等人也提出将 GAN 用于文本生成，该 GAN 模型分别将 CNN 和 LSTM 作为判别器和生成器，采取多次更新生成器后再更新一次判别器的策略来避免模式坍塌问题，同时利用矩匹配实现优化问题。

AGHAKHANI[3] 等人结合 GAN 与半监督学习提出 FakeGAN 模型用于识别众多评论网站里的虚假评论。与传统 GAN 不同，该模型采用两个分类器和一个生成器来避免模式坍塌问题。在由 800 条评论构成的数据集上进行的实验证明，该模型能达到 89.1% 的识别率。MIYATO[16] 等人提出将 GAN 中的对抗式和虚拟对抗式训练扩展应用到文本领域，通过在 RNN 的词嵌入层进行干扰以实现文本分类。实验证明该方法还可以用于解决其他文本任务，如机器翻译 [17]、学习分词的表达方式 [18] 和问答任务等，也可以用于语音或者视频生成。

2 GAN 在网络攻击检测领域的应用

作为深度学习与人工智能领域技术热点的 GAN，其“生成”和“对抗”的思想是它与网络安全之间联系的纽带，该纽带也使其在网络安全领域中展现出巨大的应用潜力和发展前景。2016 年以来，GAN 被应用于网络攻击检测，帮助检测恶意软件、恶意代码和修复漏洞 [19]，提升安全工具和反病毒软件的性能。本章从网络攻击样本生成和网络攻击行为检测两个方面介绍 GAN 在网络攻击检测领域的应用。

2.1 网络攻击样本生成

统计分析和机器学习是目前主流的恶意检测方法。虽然确实能有效检测恶意软件、恶意代码、恶意行为等，但也存在两个局限：1）训练过程中攻击数据不足，远远少于正常数据。数据集的不平衡导致检测模型失衡，无法正确检测攻击数据或者行为。2）随着技术的发展，攻击者的攻击手段也在不断改变，所采用的攻击媒介如恶意软件、恶意代码、恶意网络流等都在不断变化，这些攻击数据不会在网络上披露，无法将它们用于模型训练，导致模型无法检测未知的攻击数据。为了解决以上两个问题，研究者们引入 GAN 生成可使用的攻击数据，增强训练数据集，提升检测模型的性能。

2.1.1 恶意域名生成

袁辰 [20] 等人针对当前基于机器学习的恶意域名检测算法缺乏训练数据，无法及时更新检测模型，也无法对新产生的 DGA 域名进行识别检测，导致检测滞后且效率低下的问题，提出了域名字符生成模型 DGA-GAN，如图 2 所示。该模型包括 4 个模块：域名编码器、生成器、判别器和域名解码器。袁辰 [20] 等人基于 ASCAII 码变量方式定义域名编码器和解码器。首先去除顶级和二级域名字符得到一个 15 维域名字符向量，编码器利用 ASCAII 码转换函数将域名字符向量转换为域名 ASCAII 码向量，并利用数据归一化将向量值映射到区间 [0,1]。解码器将归一化后的向量值还原为 ASCAII 码向量值，得到域名 ASCAII 码向量，利用逆 ASCAII 码转换函数得到域名字符向量。

生成器和判别器都由 4 层神经网络构成，即输入

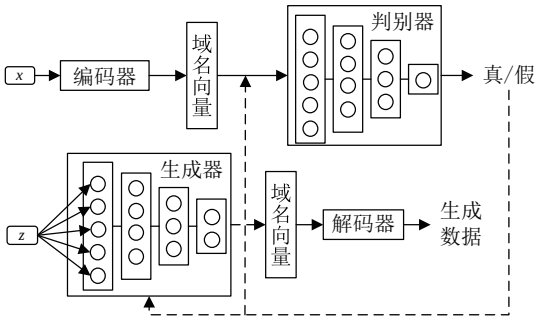

图 2 DGA–GAN 模型

层、两层隐含层和输出层。生成器接收由高斯分布模型随机产生的 100 维数据，输出 15 维域名向量；判别器接收由 15 维真实 DGA 域名向量和生成器生成的 15 维域名向量构成的 30 维域名向量，将其拆分为两个 15 维向量后进行判别。基于 Alexa 数据集和 100 万条 DGA 样本，实验从 3 个层面说明了生成数据的有效性：能否作为域名、域名的字符频率及多分类器分类检测效果对比。但模型在生成域名时并未考虑域名之间的字符上下文关系，以及该特性是否会影响生成数据的质量。

2.1.2 恶意网络流生成

为了解决训练数据缺乏所导致的检测稳定性差的问题，文献 [21] 提出了 MNF-GAN 模型。该模型利用 GAN 生成恶意网络流并验证恶意网络流是否具备可执行性和攻击性。网络流是指在网络中传输的一组数据包，一个数据包可用一个五元组表示，即 < 源 IP 地址, 目的 IP 地址, 源端口, 目的端口, 协议类型 >。文献 [21] 首先对网络流样本进行预处理，将长度统一化为 4096，每一位数据归一化到 [0,1]，即 $X=(X_1, X_2, \cdots, X_{4096})$，$X_i$ 代表样本第 i 位数据归一化后的结果；然后将样本转换为 64×64 的二维向量。

恶意网络流生成结构由生成模块、判别模块和验证模块构成，如图 3 所示。生成模块是一个多层前馈神经网络，输入包括良性样本 x 和恶意样本 m。为了保留恶意网络流的可执行性和攻击性，通过掩模操作提取出弱相关位（即改变该位的用户数据后不影响样本可执行性和攻击性的位置），在弱相关位进行扰动生成具有攻击性的对抗样本 m'。判别模块由检测器（攻击者攻击的对象）和判别器串联构成，该模块对 x 和生成的 m' 进行判别。验证模块由检测器和 Snort 网络安全检测器构成，分别验证对抗样本是否可以欺骗检测器和对抗样本是否具备攻击性。

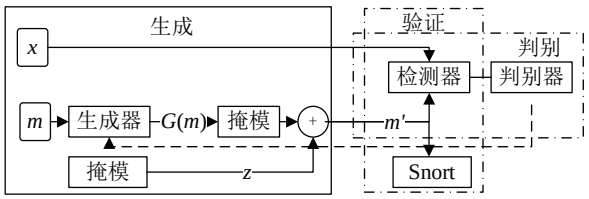

图 3 MNF–GAN 模型

基于 UNSW-NB15 数据集，实验证明，MNF-GAN 生成的恶意网络流具备可执行性和攻击性，能够有效绕过基于深度学习的网络安全检测工具。文献 [21] 通过修改恶意网络流对抗样本弱相关位的方法保留了对抗样本的可执行性和攻击性，使得生成的样本可以直接用于深度学习模型训练，具有较强的实用性。但文献 [21] 只是针对缓冲区溢出漏洞攻击网络流进行了研究，有待扩展攻击网络流的研究范围。

2.1.3 恶意软件生成

HU[22] 等人提出 MalGAN 模型用于生成恶意软件对抗样本，生成的样本能够绕过基于黑盒机器学习的检测模型。MalGAN 模型使用前向神经网络作为生成器，使用替代检测器作为判别器，以随机噪声作为输入来生成恶意样本。由于无法得知黑盒模型的内部结构，文献 [22] 首先将正常软件、恶意软件、生成的恶意软件输入待攻击的黑盒模型，得到由黑盒模型打好标签的样本；然后将得到的样本输入替代检测器，利用替代检测器来学习黑盒模型的特征；再将学习到的结果即梯度信息反馈给生成器，促使对抗样本被替代检测器检测为恶意样本的概率降低到接近为零，即实现生成器和替代检测器协同攻击黑盒模型。实验结果证明，MalGAN 模型生成的恶意软件对抗样本能够绕过黑盒检测模型。文献 [22] 的优点在于并未直接攻击黑盒模型，而是采用了替代检测器学习其梯度信息，将黑盒模型的攻击问题转换为替代检测器的样本分类问

题，并将对抗样本能否误导替代检测器的分类作为生成器好坏的评估标准。但如何使替代检测器最大限度学习黑盒模型的内部细节有待于进一步研究。

2.1.4 讨论

本节对上述 3 种恶意对抗样本生成流程通过表 1 进行了对比。总体来说，基于 GAN 的对抗样本生成主要以噪声作为输入，对抗样本和真实样本一起输入判别器进行判别，基于判别结果的反馈，生成器不断训练调整，直到判别器不可再区分真实样本和对抗样本。观察上述 3 种生成过程，生成原理和思想都是一样的，不同之处在于输入的数据根据应用场景的不同而不同，判别的步骤多少也因应用需求的不同而不同。此外，GAN 中模型的基础结构也可能不同，可以采用多层感知机、CNN、RNN 等网络结构。

表 1 生成流程比较

名称	生成流程
GAN	样本x → 判别器 → 真/假；噪音z → 生成器 → G(z)
DGA-GAN	恶意域名向量 → 判别器 → 真/假；高斯分布z → 生成器 → 生成恶意域名向量
MNF-GAN	正常网络流向量 → 检测器 → 判别器 → 正常/恶意；恶意网络流向量 → 生成器 → 生成恶意网络流向量
MalGAN	正常软件；恶意软件 → 黑盒模型 → 替代检测器 → 梯度信息；噪声z → 生成器 → 生成恶意软件

2.2 网络攻击行为检测

2.2.1 恶意代码检测

1）t-GAN 模型

传统的恶意代码检测 [23] 需要收集一定量的数据，从大量的数据中挖掘使用模式，基于模式匹配检测恶意代码。但这种方法是基于已有的数据进行检测，导致无法及时检测未知的恶意代码。为了解决这个问题，KIM[24] 等人使用迁移学习方法 [25]，将自编码器用于 GAN 模型，提出 t-GAN 模型用于检测恶意代码。

在 t-GAN 模型的训练过程中，文献 [24] 首先将恶意二进制代码转变为矩阵，再转变为恶意图像。利用二进制代码长度计算出矩阵的行数和列数，矩阵中每个数值对应一个字节，每个字节范围在 00~FF 之间，刚好对应图像的像素范围 0~255。归一化矩阵的每个数值在 0 到 1 之间。越接近 1，图像点越红；越接近 0，图像点越蓝。

基于得到的恶意图像训练自编码器的编码器和解码器，将解码器迁移为生成器。基于得到的恶意图像和生成器生成的恶意图像训练判别器，将判别器迁移为检测器用于恶意代码的检测。

基于 Kaggle Microsoft Malware Classification Challenge 数据集 [26]，文献 [24] 证明了 t-GAN 模型在精确度方面远优于传统的机器学习算法，如 *k* 邻近、朴素贝叶斯、随机森林以及决策树。但模型中 GAN 训练过程的不稳定影响了检测性能，同时模型在检测时遗漏了对 0-day 攻击的检测。

2）t-DCGAN 模型

为了提升 t-GAN 模型训练过程的稳定性，并使模型能够检测 0-day 攻击，KIM[27] 等人进一步提出了 t-DCGAN 模型。该模型使用 DCGAN 替代 t-GAN 模型中的 GAN，以提升训练过程的稳定性。KIM[27] 等人发现，恶意软件在结构上是相似的，而当以 8∶2 的比例组合两张恶意代码图像时会产生噪声，他们假设在现有恶意软件中加入噪声可生成能进行 0-day 攻击的恶意软件，基于此原理生成恶意代码图像并将其用于 t-DCGAN 模型训练。

基于 Kaggle Microsoft Malware Classification Challenge 数据集，实验将 t-DCGAN 模型与传统的机器学习算法（如 *k* 邻近、朴素贝叶斯、随机森林和决策树）以及深度学习算法（如 MLP、CNN）进行了对比。此外，还与 GAN、DCGAN 和 t-GAN 等模型进行了对比。实验结果表明，t-DCGAN 模型不仅在精确度上优于传统的机器学习算法和深度学习算法，还能有效检测 0-day 攻击。但是，t-DCGAN 模型生成的恶意图像虽然具备恶意特性，但文献 [27] 并未考虑由恶意图像转换所得到的恶意代码是否依然具备可执行性和攻击性。

2.2.2 僵尸网络检测

僵尸网络 [28] 是感染相同病毒的主机所构成的网络，该类网络严重危害了网络安全，因此僵尸网络的检测一直是一个研究热点。现有检测器不考虑网络流负载信息，如数据包长度、字节长度，只对网络流五元组，即 < 源 IP 地址，目的 IP 地址，源端口，目的端口，协议类型 > 进行分析。虽然降低了工作复杂度，但误报率较高，且无法检测新型的僵尸网络。因此，YIN[29] 等人提出 Bot-GAN 用于检测僵尸网络，其网络结构如图 4 所示。为了保留原始检测器的特点，判别器由僵尸网络检测器代替，包含一个 4 层的神经网络；生成器由一个 3 层的 LSTM 网络构成。与二分类不同，Bot-GAN 判别器输出的是三分类，即正常、异常、虚假。

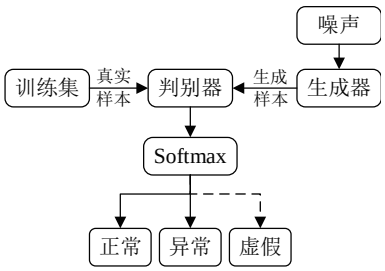

图 4 Bot–GAN 网络结构

在预处理阶段，Bot-GAN 分析不同类型僵尸网络流及其负载信息，抽取选择包括协议、字节长在内的 16 个特征；将其向量化为 122 维特征向量，其中 107 维代表协议，15 维代表其余 15 个特征；将特征向量的每一维归一化到 [0,1]，得到僵尸网络流向量。在检测阶段，将僵尸网络流向量输入判别器进行检测，并输出分类结果。

基于 ISCX 僵尸网络数据集 [30]，实验从精确率、召回率、*F*1 值、混淆矩阵 4 个方面证明了 Bot-GAN 优于现有检测器，在提升检测性能的同时降低了误报率，同时还能检测出未知僵尸网络。值得注意的是，文献 [29] 所提出的方法具有普遍性，不仅可用于检测僵尸网络，还可迁移到其他领域用于 XSS 检测等问题。

2.2.3 欺诈检测

现有在线支付欺诈检测方法主要是通过数据挖掘从大量交易日志中分析挖掘出可疑模式，基于模式利用机器学习算法进行检测分类。但相对正常交易数据而言，可用违法交易记录数据量过少，导致训练数据集失衡，降低了模型检测的有效性。CF-GAN 模型 [31] 利用 GAN 生成违法交易记录数据，综合原有违法交易记录数据构成增强数据集，用于训练分类器。

实验基于 Credit-Card 数据集 [32]，对比基于原有数据集训练的分类器，发现基于增强数据集训练的分类器在稳定性和有效性等方面都更加优越。文献 [31] 的方法优点在于可移植性强，局限在于依赖于标注数据，即训练之前需要知道数据的类别（合法或违法），难以应用到无监督学习中，也无法及时发现新型诈骗。此外，实验发现，分类器敏感性提高的同时误报率也在提升，这也是文献 [31] 未来改进方向之一。

文献 [33] 基于对抗式深度降噪自编码器（DAE）和高斯混合模型（GMM）构建 GAN-DAE 模型，有效识别受害者向欺诈者发送的转账，提高了电信诈骗的检测率。与现有检测方法相比，该方法无需依赖大量交易数据。GAN-DAE 模型结构包括 3 个模块，如图 5 所示：1）判别器。由自编码器的编码器和 GMM_1 构成。判别输入是正常转账还是虚假转账，虚假转账包括真实欺诈转账记录和由生成器生成的欺诈转账记录。判别具体过程如下：输入向量 $\boldsymbol{x}$，编码器将其转换成向量 $\boldsymbol{z}$，GMM_1 计算 $\boldsymbol{z}$ 的频率 $\Phi_1(\boldsymbol{z})$，即转账为真实转账（包括正常转账和真实欺诈转账）的概率。文献 [33] 将概率阈值设置为 0.5。2）生成器。由自编码器的解码器构成。将 $\boldsymbol{z}$ 和类别标签向量（正常、欺诈、未知）作为输入，输出欺诈转账记录的特征向量 $\boldsymbol{x}'$（文献 [33] 中设置为 168 维，包括转账金额、转账时间等特征）。3）分类器。由自编码器的编码器和 GMM_2 构成。GMM_2 计算 $\Phi_2(\boldsymbol{z})$，即转账为正常转账的概率。

文献 [33] 在由 33 万条转账记录（329820 条正常转账和 180 条虚假转账）组成的数据集上进行了实验，发现 GAN-DAE 无论是在精确率、召回率还是在错误分类率（即将正常转账误判为虚假转账的概率）上都

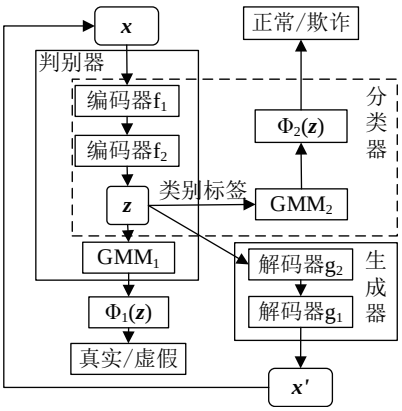

图 5 GAN–DAE 模型结构

优于现有检测方法，证明了GAN-DAE 检测的有效性。该方法优点在于无需大量训练数据，但阈值设定的合理与否影响检测的准确度。此外，该方法依赖于具有类别标签的数据，限制了其应用范围，如不能应用到无监督学习中。

2.2.4 小结

本节从特征选择、训练数据有无标签、检测结果输出、应用领域 4 个方面对上述 5 个基于 GAN 的网络攻击行为检测模型进行了对比，如表 2 所示。特征选择是指待检测的数据是否需要人工提取特征以备检测；训练数据有无标签是指训练数据是否依赖于有标签的数据；检测结果输出是指输出结果是二分类还是多分类；应用领域是指检测方法能否应用到无监督学习中。

表 2 基于 GAN 的网络攻击行为检测模型对比

项目 / 模型	特征选择	训练数据有无标签	检测结果输出	应用领域
t-GAN	无需人工选取	无	二分类（正常 / 恶意）	无监督学习
t-DCGAN	无需人工选取	无	二分类（正常 / 恶意）	无监督学习
Bot-GAN	人工选取	无	三分类（正常 / 异常 / 虚假）	无监督学习
CF-GAN	无需人工选取	有	二分类（合法 / 违法）	有监督学习
GAN-DAE	无需人工选取	有	三分类（正常 / 欺诈 / 未知）	有监督学习

从表 2 可发现，除 Bot-GAN 以外的所有模型都无需人工干预特征选择过程，但 CF-GAN 和 GAN-DAE 属于有监督学习，需要预先设定阈值，阈值设定的好坏直接影响模型的检测效果。t-GAN 是在恶意代码检测问题上 GAN 的首次应用，从精确度来看优于传统的机器学习方法，但其训练过程不稳定。为了解决该问题，提出了t-DCGAN，大大提升了训练的稳定性和检测结果的精确度。

3 结束语

本文综述了生成式对抗网络（GAN）的理论和应用研究进展，具体探究了GAN 在网络攻击检测领域的应用现状。GAN 强大的生成能力不仅使其在图像和视觉领域、语音领域、文本生成领域获得青睐，还使其在网络攻击检测领域展现出极大的应用潜力和价值。GAN 作为一个有效的深度生成式网络，可以将其纳入网络安全研究体系，促进其自身和网络安全领域的理论和应用发展。但如何生成多样化的攻击数据，以及如何解决样本长度不一对模型训练的影响，是将 GAN 应用到网络安全领域亟需解决的问题。（责编 马珂）

参考文献：

[1]GOODFELLOW I. NIPS 2016 Tutorial: Generative Adversarial Networks[EB/OL]. https://arxiv.org/abs/1406.2661,2014-6-10.

[2] ARJOVSKY M, BOTTOU L. Towards Principled Methods for Training Generative Adversarial Networks[EB/OL]. https://arxiv.org/abs/1701.04862,2017-1-17.

[3]AGHAKHANI H, MACHIRY A, NILIZADEH S, et al. Detecting Deceptive Reviews Using Generative Adversarial Networks[EB/OL].https://arxiv.org/abs/1805.10364,2018-5-25.

[4] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative Adversarial Nets[C]//NIPS. The 2014 Conference on Advances in Neural Information Processing Systems 27, December 8-13, 2014, Montreal, Canada. NY:Curran Associates, 2014. 2672-2680.

[5] SALIMANS T, GOODFELLOW I, ZAREMBA W, et al. Improved Techniques for Training Gans[EB/OL]. https://arxiv.org/abs/1606.03498,2016-6-10.

[6]ZHAO Junbo, MATHIEU M, LECUN Y. Energy-based Generative Adversarial Network[EB/OL]. https://arxiv.org/abs/1609.03126,2017-3-6.

[7]ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein Gan[EB/OL]. https://arxiv.org/abs/1701.07875,2017-12-6.

[8]GULRAJANI I, AHMED F, ARJOVSKY M, et al. Improved Training of Wasserstein Gans[EB/OL].https://arxiv.org/abs/1704.00028,2017-12-25.

[9]LEDIG C, THEIS L, HUSZÁR F, et al. Photo-realistic Single Image Super-resolution Using a Generative Adversarial Network[EB/OL]. https://arxiv.org/abs/1609.04802,2017-5-25.

[10]WU Huikai, ZHENG Shuai, ZHANG Junge, et al. GP-GAN: Towards Realistic High-Resolution Image Blending[EB/OL].

https://arxiv.org/abs/1703.07195,2017-3-25 .
[11]PATHAK D, KRAHENBUHL P, DONAHUE J, et al. Context Encoders: Feature Learning by Inpainting[C]//IEEE. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 27-30,2016, Las Vegas,USA.NJ:IEEE,2016: 2536-2544.
[12]MOGREN O. C-RNN-GAN: Continuous Recurrent Neural Networks with Adversarial Training[EB/OL]. https://arxiv.org/abs/1611.09904,2016-11-29.
[13]YU Lantao, ZHANG Weinan, WANG Jun, et al. SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient[C]//AAAI. The Thirty-first AAAI Conference on Artificial Intelligence and the Twenty-ninth Innovative Applications of Artificial Intelligence Conference, February 4-9,2017,San Francisco,California,USA. CA:AAAI, 2017: 2852-2858.
[14]LI Jiwei, MONROE W, SHI Tianlin, et al. Adversarial Learning for Neural Dialogue Generation[EB/OL]. https://arxiv.org/abs/1701.06547,2017-9-24.
[15]ZHANG Yizhe, GAN Zhe, CARIN L. Generating Text via Adversarial Training[EB/OL]. https://zhegan27.github.io/Papers/textGAN_nips2016_workshop.pdf?tdsourcetag=s_pcqq_aiomsg，2018-6-11.
[16]MIYATO T, DAI A M, GOODFELLOW I. Adversarial Training Methods for Semi-Supervised Text Classification[EB/OL]. https://arxiv.org/abs/1605.07725,2017-5-6.
[17]SUTSKEVER I, VINYALS O, LE Q V. Sequence to Sequence Learning with Neural Networks[C]//NIPS. Advances in Neural Information Processing Systems, December 8-13,2014,Montreal,Quebec,Canada.NY:Curran Associates,2016: 3104-3112.
[18]MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed Representations of Words and Phrases and Their Compositionality[C]//NIPS.Advances in Neural Information Processing Systems,December 5-10,2013,Lake Tahoe,Nevada,USA. NY:Curran Associates,2013: 3111-3119.
[19]HARER J, OZDEMIR O, LAZOVICH T, et al. Learning to Repair Software Vulnerabilities with Generative Adversarial Networks[EB/OL].https://arxiv.org/pdf/1805.07475.pdf,2018-5-28.
[20]YUAN Chen, QIAN Liping, ZHANG Hui, et al. Generation of Malicious Domain Training Data Based on Generative Adversarial Network[EB/OL].http://www.arocmag.com/article/02-2019-05-042.html,2018-3-14.
袁辰，钱丽萍，张慧，等. 基于生成对抗网络的恶意域名训练数据生成 [EB/OL].http://www.arocmag.com/article/02-2019-05-042.html,2018-3-14.
[21] PAN Yiming, LIN Jiajun.Generation and Verification of Malicious Network Flow Based on Generative Adversarial Networks[EB/OL].https://doi.org/10.14135/j.cnki.1006-3080.20180313003,2018-6-11.
潘一鸣，林家骏. 基于生成对抗网络的恶意网络流生成及验证 [EB/OL].https://doi.org/10.14135/j.cnki.1006-3080.20180313003,2018-6-11.
[22]HU Weiwei, TAN Ying. Generating Adversarial Malware Examples for Black-Box Attacks Based on GAN[EB/OL]. https://arxiv.org/abs/1702.05983,2017-2-20.
[23] ZHENG Shengjun, GUO Longhua, CHEN Jian, et al. An Online Detection System for Advanced Malware Based on Virtual Execution Technology [J]. Netinfo Security, 2016,16 (1) : 29-33.
郑生军，郭龙华，陈建，等. 基于虚拟执行技术的高级恶意软件攻击在线检测系统 [J]. 信息网络安全，2016,16 (1) : 29-33.
[24]KIM J Y, BU S J, CHO S B. Malware Detection Using Deep Transferred Generative Adversarial Networks[C]//ICONIP. the 24th International Conference on Neural Information Processing, November 14-18,2017,Guangzhou,China.HK:Springer,2017: 556-564.
[25]ARNOLD A, NALLAPATI R, COHEN W W. A Comparative Study of Methods for Transductive Transfer Learning[C]//IEEE. ICDM Workshops 2007,October 28,2007,Omaha,USA.NJ:IEEE, 2007: 77-82.
[26]MICROSOFT. Microsoft Malware Classification Challenge (BIG 2015)[EB/OL].https://www.kaggle.com/c/malware-classification.2015-2-1.
[27]KIM J Y, BU S J, CHO S B. Zero-day Malware Detection Using Transferred Generative Adversarial Networks Based on Deep Autoencoders[J]. Information Sciences, 2018, 460(9): 83-102.
[28]HEN Hongsong, WANG Gang, SONG Jianlin. Research on Anomaly Behavior Classification Algorithm of Internal Network User Based on Cloud Computing Intrusion Detection Data Set[J]. Netinfo Security, 2018 ,18(3) : 1-7.
陈红松，王钢，宋建林. 基于云计算入侵检测数据集的内网用户异常行为分类算法研究 [J]. 信息网络安全，2018，18 (3) : 1-7.
[29] YIN Chuanlong, ZHU Yuefei, LIU Shengli, et al. An Enhancing Framework for Botnet Detection Using Generative Adversarial Networks[C]//IEEE.2018 International Conference on Artificial Intelligence and Big Data (ICAIBD),May 26-28,2018,Chengdu,China.NJ:IEEE, 2018:228-234.
[30]SHIRAVI A, SHIRAVI H, TAVALLAEE M, et al. Toward Developing a Systematic Approach to Generate Benchmark Datasets for Intrusion Detection[J]. Computers & Security, 2012, 31(3): 357-374.
[31]FIORE U, DE S A, PERLA F, et al. Using Generative Adversarial Networks for Improving Classification Effectiveness in Credit Card Fraud Detection[EB/OL]. https://www.sciencedirect.com/science/article/pii/S0020025517311519,2017-12-30.
[32]DAL P A, CAELEN O, JOHNSON R A, et al. Calibrating Probability with Undersampling for Unbalanced Classification[C]//IEEE. 2015 IEEE Symposium Series on Computational Intelligence, December 7-10,2015,Cape Town,South Africa.NJ:IEEE,2015: 159-166.
[33] ZHENG Yujun, ZHOU Xiaohan, SHENG Weiguo, et al. Generative Adversarial Network-based Telecom Fraud Detection at the Receiving Bank[J]. Neural Networks, 2018, 102(6): 78-86.