基于生成对抗网络的僵尸网络检测

邹福泰，谭越，王林，蒋永康
（上海交通大学网络空间安全学院，上海 200240）

摘 要：为了解决僵尸网络隐蔽性强、难以识别等问题，提高僵尸网络检测精度，提出了基于生成对抗网络的僵尸网络检测方法。首先，通过将僵尸网络流量中的数据包重组为流，分别提取时间维度的流量统计特征和空间维度的流量图像特征；然后，基于生成对抗网络的僵尸网络流量特征生成算法，在 2 个维度生产僵尸网络特征样本；最后，结合深度学习在僵尸网络检测场景下的应用，提出了基于 DCGAN 的僵尸网络检测模型和基于 BiLSTM-GAN 的僵尸网络检测模型。实验表明，所提模型提高了僵尸网络检测能力和泛化能力。

关键词：僵尸网络；深度学习；流量分析；机器学习；生成对抗网络
中图分类号：TP393.08
文献标识码：A
DOI: 10.11959/j.issn.1000−436x.2021082

Botnet detection based on generative adversarial network

ZOU Futai, TAN Yue, WANG Lin, JIANG Yongkang
School of Cyber Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

Abstract: In order to solve the problems of botnets' strong concealment and difficulty in identification, and improve the detection accuracy of botnets, a botnet detection method based on generative adversarial networks was proposed. By reorganizing the data packets in the botnet traffic into streams, the traffic statistics characteristics in the time dimension and the traffic image characteristics in the space dimension were extracted respectively. Then with the botnet traffic feature generation algorithm based on generative adversarial network, botnet feature samples were produced in the two dimensions. Finally combined with the application of deep learning in botnet detection scenarios, a botnet detection model based on DCGAN and a botnet detection model based on BiLSTM-GAN were proposed. Experiments show that the proposed model improves the botnet detection ability and generalization ability.

Keywords: botnet, deep learning, traffic analysis, machine learning, GAN

1 引言

随着全球网络威胁态势的不断变化，新型网络攻击手段层出不穷，攻击方式愈发复杂多变，网络安全态势依旧严峻，网络安全感知防护仍然任重道远。在木马、蠕虫等恶意软件之中，僵尸网络（botnet）仍然是最广泛、最持久的安全威胁。根据 CenturyLink 威胁研究实验室发布的《2019 威胁报告》[1]，2019 年上半年，某实验室的系统平均每天检测到恶意威胁 120 万种，较 2017 年的近 20 万种增加了 500%。其中，僵尸网络威胁达到了每日约 18 000 多次。

僵尸网络指的是一组被攻击者破坏了安全防护系统并夺取了控制权的用户终端。受感染终端（bot）通常由一个或多个攻击者进行控制[2]。通过僵尸网络，攻击者可以实施一系列的犯罪活动，包括分布式拒绝服务（DDoS, distributed denial of service）攻击、网络钓鱼、加密勒索、恶意软件分

收稿日期：2020−09−25；**修回日期**：2020−12−20
基金项目：国家重点研发计划基金资助项目（No.2020YFB1807500）
Foundation Item: The National Key Research and Development Program of China (No.2020YFB1807500)

发以及交换非法信息等，对互联网生态环境及网络空间安全构成了严重威胁。

近年来，已有多起僵尸网络攻击导致的安全事件发生。其中，2016 年年底，Mirai 僵尸网络利用嵌入式和物联网设备进行了大规模的 DDoS 攻击，并在最初的 20 小时内感染了近 65 000 个物联网设备[3]。2017 年，Necurs[4-5]僵尸网络几小时内发送了超过 40 000 封恶意邮件，向数以千计的受害者勒索 2.047 BTC。2018 年，门罗币挖矿僵尸网络 Smominru[6]感染全球 3 000 万个系统，并破坏 50 万台用于加密采矿的机器。2019 年，Emotet 僵尸网络通过电子邮件中的下载链接，控制受害者终端并窃取终端数据，感染美国多个州和地方政府的系统[7]。

Destil 安全研究实验室发布的《2019 恶意僵尸报告》显示[8]，2019 年僵尸网络流量在全球互联网流量总占比近 40%。此外随着物联网的兴起，僵尸网络越来越多地利用隐私结构和匿名服务。麻省理工互联网研究计划（IPRI, Internet policy research initiative）[9]报告显示，Skynet 僵尸网络、Sefnet 僵尸网络和 Zeus 僵尸网络变体等使用 Tor 匿名网络托管的僵尸网络正在肆虐，为检测定位真实 IP 地址增加了难度。

因此，从网络空间安全威胁态势来看，僵尸网络攻击手段越来越先进，越来越难以被发现，网络空间安全形式越来越严峻，有必要开发更加新颖、更加有效的僵尸网络检测技术，更好地防范僵尸网络对网络空间环境的影响。

本文基于对僵尸网络流量的研究，提出利用生成对抗网络（GAN, generative adversarial network）从时间、空间 2 个维度来检测僵尸网络的方法。本文提出的基于生成对抗网络的僵尸网络检测增强模型，可以提高检测模型的检测能力及泛化能力，并提供了全面有效地利用深度学习检测僵尸网络的方案，对僵尸网络检测和网络安全管理具有重要意义。具体而言，本文的主要贡献如下。

1) 针对部分僵尸网络训练样本较少的问题，本文研究了生成对抗网络在图像识别领域的建模应用，提出了一种基于 DCGAN（deep convolutional generative adversarial network）的僵尸网络检测模型，可以有效扩充标签样本集，提高检测模型的检测能力，增强模型的泛化能力，为进一步提升僵尸网络检测与发现能力提供了一种有效的建模方法。

2) 提出了一种基于 BiLSTM 的僵尸网络时序特征生成器建模方法 BiLSTM-GAN，并实现了一种新的基于 BiLSTM-GAN 的僵尸网络检测模型，有效提高了对僵尸网络行为的检测准确率，提升了对未知僵尸网络的检测能力。

2 技术背景

随着云服务和物联网（IoT, Internet of things）设备的数量呈指数增长，僵尸网络攻击事件大大增加，僵尸网络安全防护迫在眉睫。McAfee Labs 报告[10]显示 2019 年第一季度发现的新恶意软件数量达到 6 600 万，其中僵尸网络占了很大一部分。传统的基于人工的可疑代码检测、逆向工程和漏洞识别的方法耗时较高，以至于无法满足日益增长的高精度、高实时性的僵尸网络检测能力要求。由于僵尸网络近年来的演化，其种类越发繁多、破坏力不断增强、对网络安全造成的威胁日益严峻，国内外现已对僵尸网络形成了有针对性的研究。目前在此领域下已经产生了许多对僵尸网络的不同检测方法，按照检测算法及数据来源，可将其分为两大类：基于蜜罐的僵尸网络检测方法、基于入侵检测系统（IDS, intrusion detection system）的检测方法。其中基于入侵检测系统的检测方法分类如图 1 所示。

在现有的检测方法中，最常见的是基于异常的检测方法。其原理是通过利用机器学习、统计分析等方法对网络中与预期行为不一致的异常数据进行检测。

基于异常的僵尸网络检测方法依据检测对象可分为基于主机的检测方法和基于网络的检测方法，以及主机和网络相结合的检测方法[11]。基于主机的检测方法针对受感染计算机上运行的僵尸恶意软件，主要通过检查终端级别的主机信息，如 API（application program interface）调用、文件更改、活动进程、资源使用情况等来检测终端是否被感染。Tokhtabayev 等[12]提出了一种监控终端系统调用的监视网络检测方法。Sharafaldin 等[13]提出了一种使用内存取证分析技术和新型的域生成算法检测器来检测并可视化僵尸网络的方法。Creech 等[14]提出了一种基于连续和非连续的系统调用的语义分析僵尸网络检测方法。

基于网络的检测方法基于对网络流量的分析[15]，通过分析僵尸网络生命周期不同阶段产生的网络

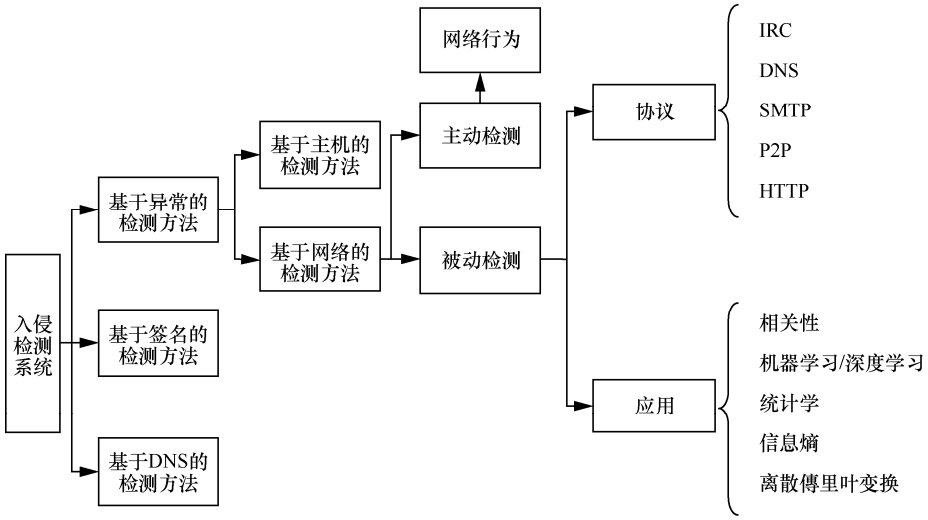


图 1　基于入侵检测系统的检测方法分类

流量的不同参数（包括网络流量行为、流量模式、响应时间、网络负载和连接特征等）来感知恶意流量。基于网络的检测方法可以进一步分为 2 类，即主动检测和被动检测。

基于网络的主动检测方法通过向被监视的网络或服务器中注入特制的数据包，并对捕获的响应进行观察和分析，以检测网络中的恶意活动。Gu[16]提出了一种主动检测技术 BotProbe 来检测基于 IRC 的僵尸网络，它通过注入调查终端内部的数据包来测试 IRC 终端是否是僵尸。Yahyazadeh 等[17]提出了一种基于网络行为的主动检测方法 BotCatch，该方法通过在线增量聚类算法识别参与协作活动的可疑主机，然后基于几个模糊函数为每个主机计算一个分数，以识别被僵尸感染的主机。但基于网络的主动检测方法也有局限性：其会向可疑终端注入额外的数据包，使正常的网络流量过载；难以将合法流量与人工注入的流量分开以进行异常检测，干扰正常流量并有可能存在隐私泄露问题。

基于网络的被动检测方法通过观察网络流量以发现僵尸程序或 C&C 服务器发起的任何可疑通信。这种方法的基本思路是：存在于同一僵尸网络中的僵尸程序应该具有相同的通信模式，可以通过在某些指定时间段内观察流量的请求–响应行为来检测。目前，研究人员已经提出了许多基于网络的被动僵尸网络检测技术，如 Gu 等提出了 BotHunter[18]、BotSniffer[19]和 BotMiner[20]，其中 BotHunter 利用僵尸感染阶段的有序通信流量进行检测，BotSniffer 和 BotMiner 则是利用属于同一僵尸网络的僵尸具有相似的活动和响应进行检测。Zhao 等[21]通过分析数据流行为特征和流间特征建立决策树分类模型，有效检测僵尸网络。此外，基于网络的被动检测方法所使用的具体应用模型又可分为统计方法、图论、机器学习、相关性、熵、随机模型、离散时间序列、傅里叶变换、数据挖掘、聚类分析、可视化等技术，以及这些技术的结合[22]。

随着人工智能的发展，神经网络深度学习算法逐渐被应用于僵尸网络检测。Torres 等[23]将僵尸网络流量转化成状态特征序列，并采用循环神经网络（RNN, recurrent neural network）进行检测。Homayoun 等[24]采用自编码器（AE, auto-encoder）和卷积神经网络（CNN, convolutional neural network）这 2 种深度学习算法检测恶意僵尸网络流量。Vinayakumar 等[25]针对恶意的域名生成算法（DGA, domain generation algorithm）提出了一种可扩展的结合了卷积神经网络与长短时记忆网络的检测模型，可以有效检测基于 DGA 的僵尸网络。Mcdermott 等[26]将深度学习算法应用到物联网领域，建立了一种基于双向长短时记忆的递归神经网络，通过文本识别检测物联网僵尸网络。Meidan 等[27]同样研究了物联网僵尸，提出了一种使用深度自动编码器检测物联网僵尸异常流量的方法，并对 Mirai 和 Bashlite 这 2 种知名物联网僵尸网络进行了实验评估，取得了不错的效果。

在僵尸网络领域，深度学习有以下优点：易于实现特征工程、易于实现高维空间特征表达、易于实现大数据学习。

目前，深度学习在僵尸网络领域的应用主要可

以分为 2 个方向[28]，第一个方向是利用深度学习模型中的生成模型，对高维特征逐层转换，学习其抽象特征，完成对僵尸网络的检测分类；第二个方向是采用深度学习模型中的判别模型自动学习僵尸网络的内在特征，并通过 Softmax 等函数直接进行判别分类。

本文提出了一种利用生成对抗网络，从时间和空间 2 个维度检测僵尸网络的方法。该方法首先将僵尸网络流量中的数据包重组为流，分别提取时间维度的流量统计特征和空间维度的流量图像特征；然后基于生成对抗网络的僵尸网络流量特征生成算法，在 2 个维度生产僵尸网络特征样本；最后设计基于 DCGAN 的僵尸网络检测模型和基于 BiLSTM-GAN 的僵尸网络检测模型，实现基于深度学习的僵尸网络流量检测方法。

3 本文方法介绍

3.1 生成对抗网络

为了提高检测模型的检测能力和稳健性，本文在传统机器学习方法上应用生成对抗网络，从空间和时间 2 个维度分别研究僵尸网络特征的生成对抗。

GAN 近年来在深度学习领域受到广泛关注，其出众的数据生成能力已成为深度学习领域最重要、最热门的研究课题。GAN 主要思想源自博弈论，通过设置判别模型与生成模型来互相竞争，以提升学习的效果，获得高维、复杂的真实样本数据分布。其结构如图 2 所示。

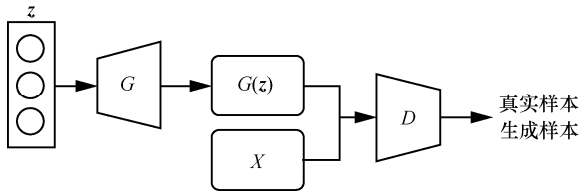


图 2 GAN 结构

GAN 的主要结构包含一个生成器和一个判别器，生成器 G 输入随机噪声矢量 $\boldsymbol{z}$（通常为均匀分布或正态分布），通过将其映射至新的多维数据空间，生成了假样本 $G(\boldsymbol{z})$；然后使用判别模型进行二分类以计算测试样本为真实样本的可能性并与真实情况进行比较。当判别器 D 的判别准确率达到 50% 时，意味着已经无法确定测试样本是否为真实样本或生成的假样本，这时认为生成器 G 已经学习到了真实样本的数据分布。生成对抗网络目标函数为

$$\min_G \max_D V(D,G) = E_{x\sim P_{\text{data}}(x)}[\log D(x)] + E_{z\sim P_z(z)}[\log(1-D(G(z)))] \quad (1)$$

其中，$\boldsymbol{x}$ 表示数据样本矩阵，$P_z(\boldsymbol{z})$表示输入噪声的数据分布，$G(\boldsymbol{z})$表示输入噪声生成的样本数据，$D(\boldsymbol{x})$表示判别器所判断的样本 $\boldsymbol{x}$ 为真实样本而非生成样本的概率。

目前，生成对抗网络在许多领域已经得到了充分的应用，尤其在计算机视觉领域取得了很大的突破，主要被应用于图像生成、图像到图像翻译、提高图像分辨率和图像补全等领域。

在网络安全领域，Kim 等[29]提出了一种基于迁移学习的生成对抗网络 tGAN，将自编码器应用于生成对抗网络，进而对恶意代码检测。此后 Kim 等[30]又进一步提出了 tDCGAN，改善了模型训练的稳定性，并使其具备了一定的检测 0-day 攻击的能力。Yin 等[31]提出了 Bot-GAN，用于僵尸网络检测，其输入为僵尸网络的统计特征，通过生成对抗网络以提高检测模型的精度。

本文研究生成对抗网络在僵尸网络领域的应用，从空间和时间 2 个维度来实现僵尸网络流量特征的生成，以增加僵尸网络流量特征样本，从而提高僵尸网络检测模型准确率和稳健性。

3.2 数据预处理

3.2.1 基于 ResNet 的僵尸网络检测模型

基于 ResNet 的僵尸网络检测模型总体框架[28]包括数据预处理、训练及测试模块。首先对 ISCX botnet 数据集进行数据预处理，将其转换成模型所需图像。数据预处理分为以下几个步骤。

1) 流量分割

僵尸网络流量分类研究的主体对象需要按照一定的粒度分割成特定的流量单元。本文采用僵尸网络领域常用的流粒度。流被定义为拥有相同五元组且按照时间顺序排列的数据包的集合。定义单个数据包为(q,l,t)，其中 q 代表该数据包的五元组，即<src_ip,src_port,dst_ip,dst_port,protocol>；l 代表该数据包的长度；t 代表该数据包的起始时间。则单条数据流可定义为

$$p=(q,l,t)$$
$$f=\{p_1=(q_1,l_1,t_1),p_2=(q_2,l_2,t_2),\cdots,p_n=(q_n,l_n,t_n)\}=(Q_f,L_f,D_f,T_f) \quad (2)$$

其中，$q_1=q_2=\cdots=q_n$；$t_1<t_2<\cdots<t_n$ 表示流中数据包按起始时间排序；Q_f 代表单条流中所有数据包

一致拥有的相同五元组；$L_f = l_1 + l_2 + \cdots + l_n$ 代表流中所有数据包的长度总和，即单条流的长度，$D_f = t_n - t_1$ 代表单条流的持续时间；$T_f = t_1$ 代表单条流的起始时间，即流中的第一个数据包的起始时间。

将流量分割为流粒度之后，还需要考虑数据包内的多个协议层信息。部分研究工作只选取应用层信息，而本文考虑到检测全面性等问题，选用全部数据信息进行流量分割处理。

2) 流量清洗

为了防止流量中的敏感信息泄露，在使用数据前需要在流量处理阶段对数据进行流量清洗以保护用户隐私。比较常用的方法是将 MAC 地址和 IP 地址分别进行随机化处理。本文在这一步骤中也对重复的数据进行清理，消除冗余。

3) 图像转化

由于分割后的数据流长度不定，而本文所使用的针对图像特征的 ResNet 检测模型需要使用相同维度的输入图像，因此需要对数据流内容进行截取处理。目前，常见的截取方法如下：① 直接截取原始数据的前 24 B，24 个数据包为一组，组成图像；② 使用 TCP 层的前 1 024 B 流量；③ 使用 MNIST 数据集的格式，并采用网络流量中的前 784 B 转化为 28 像素×28 像素的图像。本文直接提取前 1 024 B，对于长度不足 1 024 B 的，使用 0x00 填充，最后转化为 32 像素×32 像素的图像。

3.2.2 基于 BiLSTM 的僵尸网络检测模型

基于 BiLSTM 的僵尸网络检测模型总体框架包括数据预处理、训练及检测模块。首先将 ISCX botnet 数据集进行数据与处理以获得检测所需要的统计特征。数据的预处理主要分为以下几个步骤。

1) 流量分割

在本节的检测模型中，依然需要将网络流量分割为一定的流量单元，所采用的方法与 3.2.1 节的方法相同，依照五元组分割流量数据。

2) 特征提取

本文参考了现有的研究[32]，在常见的特征基础之上，筛选出部分能够反映僵尸网络流量通信和行为特点且真实有效的统计特征，进一步优化了僵尸网络统计特征的选择与提取。本节提取的流量统计特征如表 1 所示。

如表 1 所示，提取的流量特征可以分为三大类：基本特征、基于异常行为的特征和基于流相似的特征。基本特征反映了流量的协议特征，主要包括流量的源 IP、目的 IP、源端口、目的端口、通信协议等；基于异常行为的特征反映了流量的通信连接特征和连接行为特征，如流持续时间、重连接次数等，可以用于刻画僵尸网络的通信模式；基于流相似的特征反映了流的数据特征和时间特征，由于僵尸网络生成的网络流量与普通流量相比更加相似且统一，因此基于流相似的特征可以很好地刻画僵尸网络流量的特性。例如，僵尸网络在发动 DDoS 攻击时，常常会产生大量相同长度的数据包。

表 1　流量统计特征

特征名称	含义	类型
Protocol	传输层协议	基本特征
Duration	流持续时间	基于异常行为的特征
Reconnect	重连次数	基于异常行为的特征
PX	交换包数量	基于异常行为的特征
IOPR	输入包数量/输出包数量	基于异常行为的特征
NNP	交换空包数	基于异常行为的特征
NSP	交换小数据包数量	基于异常行为的特征
PSP	交换小数据包百分比	基于异常行为的特征
FPS	第一个包的长度	基于异常行为的特征
TBT	总字节数	基于流相似的特征
APL	平均数据包长度	基于流相似的特征
DPL	相同长度包数/总包数	基于流相似的特征
PV	数据包长度标准差	基于流相似的特征
BS	每秒平均比特数量	基于流相似的特征
PPS	每秒平均包数量	基于流相似的特征
AIT	数据包平均到达时间	基于流相似的特征

3) 数值化以及统一化

不同的特征可能有不同的量纲和单位，因此需要对特征数据进行数值化与归一化以消除数据类型、大小之间的差异。

本节对提取到的 16 种特征中的非数值特征 protocol 进行数值化处理，protocol 共有 107 种可能的取值，因此可以利用 One-Hot 编码将其编码成 107 维特征向量。加上 15 种数值特征，组成 122 维的特征向量。然后使用 min-max 归一化方法对得到的特征向量进行归一化，使所有维度的数据分布在[0,1]，消除量纲的影响。

3.3 基于生成对抗网络的僵尸网络检测模型

3.3.1 基于 DCGAN 的僵尸网络检测模型

随着生成对抗网络在图像生成领域研究的不断深入，出现了许多优秀的图像生成算法，但是基

本的 GAN 训练不稳定，容易出现生成器产生无意义的输出的现象。为此，Radford 等[33]提出了深度卷积生成对抗网络——DCGAN，创新地将基本 GAN 中的生成器的全连接层替换为反卷积层，从而在图像生成任务中实现了出色的性能。DCGAN 利用 CNN 强大的特征提取能力提高了生成网络的学习效果，通过在层内使用批标准化（BN, batch normalization）让生成器得以稳定学习，使模型可以更好地学习样本数据分布，更稳定地生成高质量的图片。

标准的生成对抗网络是一个二分类模型，但是对于僵尸网络检测来讲，判别器的输入样本包括真实样本中的良性流量样本和僵尸网络流量样本以及生成器生成的假样本 $G(z)$。因此，本节提出的基于 DCGAN 的僵尸网络检测模型将基于 ResNet 的僵尸网络检测模型作为判别器，并修改其输出节点为 3 个，分别对应良性样本（benign）、僵尸样本（botnet）、生成器生成样本（fake），即为三分类模型。基于 DCGAN 的僵尸网络检测模型如图 3 所示。

基于 DCGAN 的僵尸网络检测模型可以源源不断地生成网络流量图像，扩充僵尸网络训练集，并通过生成对抗网络的反馈机制提升检测模型的准确性。基于 DCGAN 的僵尸网络检测模型使用改进的交叉熵损失函数为

$$L_c = -E_{x,y\sim P_{\text{data}}(x,y)} \log P_{\text{model}}(c \mid x) \tag{3}$$

其中，(x,y) 表示所有输入样本及样本标签的集合，包括真实样本 $(x_{\text{true}}, y_{\text{true}})$ 和生成样本 $(x_{\text{false}}, y_{\text{false}})$ 两部分；c 表示样本所属的类别且 $c \in \{\text{Benign,Botnet, Fake}\}$；$P_{\text{model}}(c \mid x)$ 表示检测模型预测的样本 x 属于类别 c 的概率。

3.3.2 基于 BiLSTM-GAN 的僵尸网络检测模型

近年来，随着自然语言处理领域研究不断深入以及生成对抗网络的日渐成熟，逐渐出现了一些时间序列生成领域的成果。Oord 等[34]提出了用于生成原始音频的 WaveNet。Mehri 等[35]提出了用于语音合成的 SampleRNN。Mogren[36]提出了一种使用一个 LSTM 层和一个全连接层组成生成器和判别器的 C-RNN-GAN，用于生成古典音乐。Yu 等[37]提出了一种使用 2 个 LSTM 层和一个全连接层组成生成器和判别器的 C-LSTM-GAN，通过音乐的旋律生成歌词。在这些研究的基础上，Zhu 等[32]提出了一种使用 BiLSTM 组成生成器，CNN 组成判别器的 BiLSTM-CNN-GAN，用于生成医用心电图。在参考现有的时间序列生成算法的研究基础上，本文提出了一个基于 BiLSTM-GAN 的僵尸网络检测模型，如图 4 所示。

基于 BiLSTM-GAN 的僵尸网络检测模型同 3.3.1 节一样使用改进后的交叉熵损失函数，不断生成僵尸网络统计特征样本，提高检测模型的检测准确率。

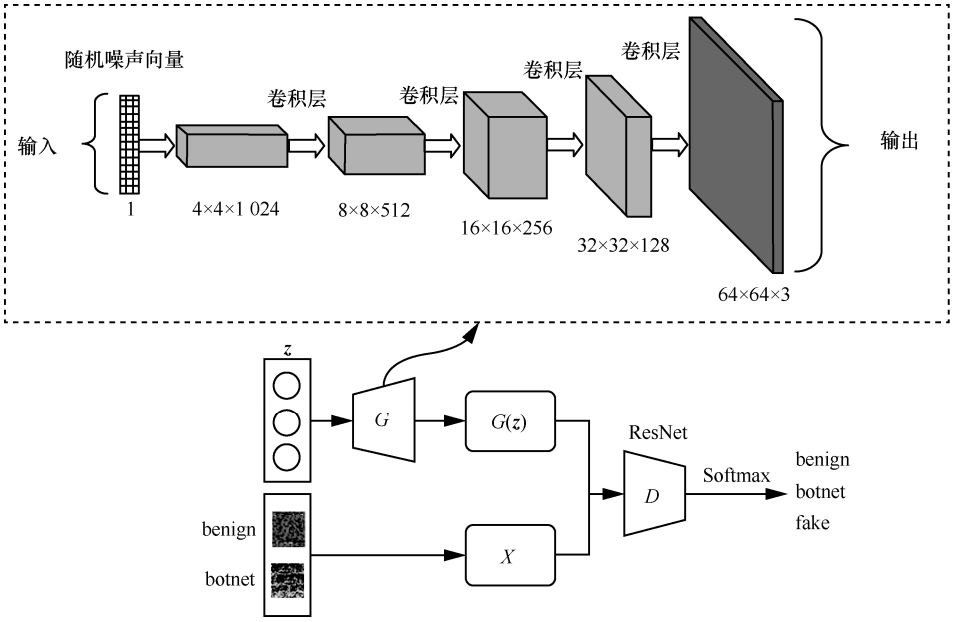


图 3 基于 DCGAN 的僵尸网络检测模型

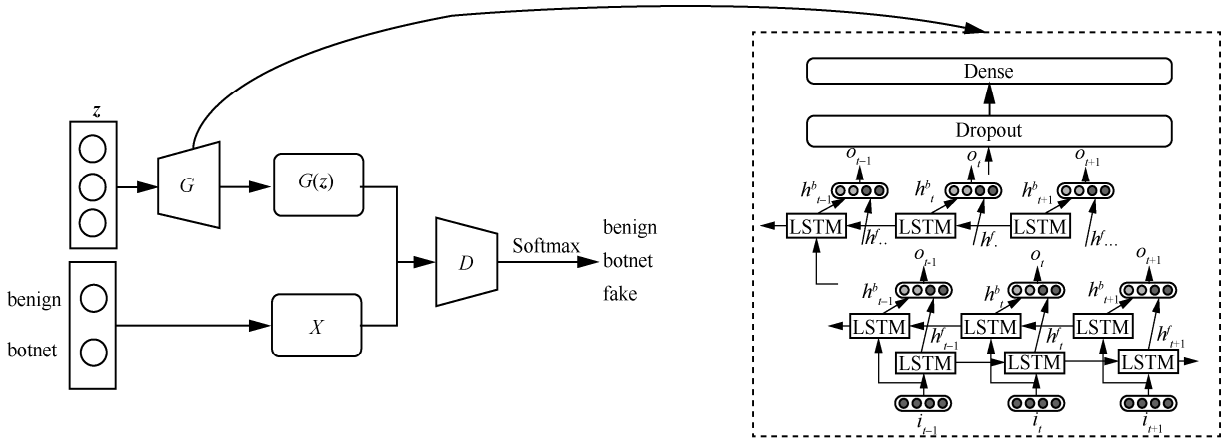


图 4　基于 BiLSTM-GAN 的僵尸网络检测模型

4　实验及分析

4.1　数据集

所有检测模型的实际检测效果都和训练及评估检测模型所采用的数据集息息相关，因此选择一个合适的、有效的数据集对于模型训练是至关重要的。

目前，僵尸网络检测领域常用的实验数据集有 CTU-13 数据集[38]、ISOT 数据集[21]、Bot-IoT 数据集[39]、ISCX2012 入侵检测数据集[40]、N_BaIoT 数据集[41]、ISCX botnet 数据集[42]等公开数据集。但目前大多数的僵尸网络数据集存在以下 3 个方面的问题。

1) 通用性较差

大多数僵尸网络数据集只包含极个别的僵尸网络流量数据，存在僵尸网络数据样本多样性较差的问题，这样训练得到的检测模型只能描述特定的僵尸网络行为的少部分特征，不具备很好的通用性，且面对新的僵尸网络威胁，无法产生良好的效果。如，ISOT 数据集仅包含了 Storm 和 Zeus 这 2 种 P2P 类型的僵尸网络；NBaIoT 数据集仅包含了 Mirai 和 BASHLITE 这 2 种物联网类型的僵尸网络。使用这些数据集实现的检测模型虽然检测精度较高，但面对实际僵尸网络流量时很难有较好的检测效果。

2) 不能很好地反映实际情况

大多数的僵尸网络数据集都是在受控环境中进行生成或捕获的。较实际的僵尸网络，受控环境中的样本很难实现所有预期的恶意行为，从而使数据集中的数据无法全面实际地展现僵尸网络特征。另外，受控环境中的样本难以长时间进行生成或捕获，会造成部分处于静默休眠期的僵尸网络采集不全面。

3) 可能存在隐私问题

大多数的僵尸网络数据集收集的网络流量为了能够反映检测模型在部署期间面对真实环境的检测能力，通常会在数据集中混入实际生活中的网络流量。但由于隐私问题，在大多数情况下，在实际生产环境中捕获网络流量是不可行的，只能在受控环境中生成或捕获流量。

为了解决以上 3 个问题，本文在评估了多个数据集后，选用了加拿大纽布伦斯威克大学网络安全研究所创建的僵尸网络数据集——ISCX botnet 数据集，作为训练和评估僵尸网络检测模型检测效果的基准数据集。该数据集通过覆盖方法[43]合并多个知名数据集，并将恶意僵尸网络流量数据映射到本地网络与良性数据合并，克服了传统僵尸网络数据集存在的问题。ISCX botnet 数据集分为训练集和测试集两部分，共包含 16 种不同类型的僵尸网络流量及大型网络中的未知良性流量。在进行流量分割后，训练集包含流数据 347 121 个，经数据处理后包含流量数据 162 410 个，其中，良性流量 141 887 个，僵尸网络流量 20 523 个；测试集包含流数据 321 893 个，经数据处理后包含流量数据 133 261 个，其中，良性流量 112 302 个，僵尸网络流量 20 959 个。训练集包含了 15%的 ISOT 数据集，ISCX2012IDS 数据集中部分良性流量和由捷克技术大学创建的恶意软件捕获项目所捕获的 CTU-13 数据集中的 Neris、Rbot、Virut 和 NSIS 这 4 种僵尸网络流量；测试集包含 25%的 ISOT 数据集，ISCX2012IDS 数据集中

部分良性流量和 IRC 僵尸网络流量和 CTU-13 数据集中的 Neris、Rbot、Virut、NSIS、Menti、Sogou 和 Murlo 这 7 种僵尸网络流量。在保证数据集更符合实际情况的基础上，测试集的僵尸网络种类多于训练集的僵尸网络种类，保证了数据集拥有评估检测模型是否具备检测未知僵尸网络的能力。ISCX botnet 数据集僵尸网络组成情况如表 2 所示，其中√和×分别表示僵尸网络样本存在和不存在，百分数表示僵尸网络样本占比。

表 2 ISCX botnet 数据集僵尸网络组成情况

僵尸网络	类型	训练集	测试集
Neris	IRC	√(12%)	√(5.67%)
Rbot	IRC	√(22%)	√(0.018%)
Menti	IRC	×	√(0.62%)
Sogou	HTTP	×	√(0.019%)
Murlo	IRC	×	√(1.06%)
Virut	HTTP	√(0.94%)	√(12.8%)
NSIS	P2P	√(2.48%)	√(0.165%)
Zenus	P2P	√(0.01%)	√(0.109%)
SMTP Spam	P2P	√(6.48%)	√(4.72%)
UDP Storm	P2P	×	√(9.63%)
Tbot	IRC	×	√(0.283%)
Zero Access	P2P	×	√(0.221%)
Weasal	P2P	×	√(9.25%)
Smoke Bot	P2P	×	√(0.017%)
Zenus control (C&C)	P2P	√(0.01%)	√(0.006%)
ISCX IRC bot	P2P	×	√(0.387%)

4.2 实验过程设计

实验环境系统配置为 CentOS7.7，CPU 为 10 核 Intel(R)Xeon(R)CPUE5-2630v4@ 2.20 GHz，采用 2 个 Nvidia 1080Ti GPU 加速训练，深度学习神经网络模型使用主流的 Keras 深度学习框架进行搭建。

在模型结构设计方面，基于 DCGAN 的检测模型输入为 3.3 节中处理后的僵尸网络流量图片以及数据采用正态分布的噪声 z。生成器由多个卷积层、上采样层和 LeakyReLU 激活层组成并进行图像样本生成。其中，卷积层采用 5×5 卷积核，上采样层采用 2×2 的上采样因子，LeakyReLU 激活层负斜率系数设为 0.2，迭代次数设为 50 次，批处理个数 batch_size 设为 128，选用 Adam 作为优化器，式(3)所示的改进后的交叉熵损失函数作为生成器的损失函数。判别器采用 3.1 节中所述的基于 ResNet 的僵尸网络检测模型，输出节点设为 3 个，分别表示模型预测输入的网络流量是良性流量、僵尸网络流量还是生成器生成的流量。

为了进行对比实验，本节还实现了基于对抗自编码器（AAE, adversarial autoencoder）的检测模型和基于基础 GAN 的检测模型。

基于 BiLSTM-GAN 的检测模型输入为 3.2 节中处理后的僵尸网络流量统计特征以及数据采用均匀分布的噪声 z。生成器由四层架构组成，包括一个输入层，一个输出层和 2 个 BiLSTM 隐层，进行僵尸网络时间特征生成。其中，输入层节点设为 128 个，BiLSTM 隐层节点为 64 个，输出层节点为僵尸网络流量统计特征维数即 122 个，迭代次数设为 100 次，批处理个数 batch_size 设为 128，选用 Adam 作为优化器。判别器采用基于 BiLSTM 的僵尸网络检测模型，输出节点设为 3 个。

4.3 实验结果评估

本文在全为已知僵尸网络的情况下对基于 GAN、DCGAN、AAE 的检测模型分别混入 100、500、1 000、2 000、5 000、8 000 个生成器生成的样本，观测 3 种检测模型和 3.2 节中所述的基于 ResNet 的僵尸网络检测模型的检测性能指标。

通过分析生成器生成的样本可知，相较于 GAN，DCGAN 学习到的数据分布更详细，样本之间的差异也更为明显。AAE 通过调整输入噪声 z，避开了 GAN 无法生成离散样本的问题，使样本更平滑，但也具有自编码器存在的分辨率较低的问题。

4 种僵尸网络检测模型的检测准确率如图 5 所示。

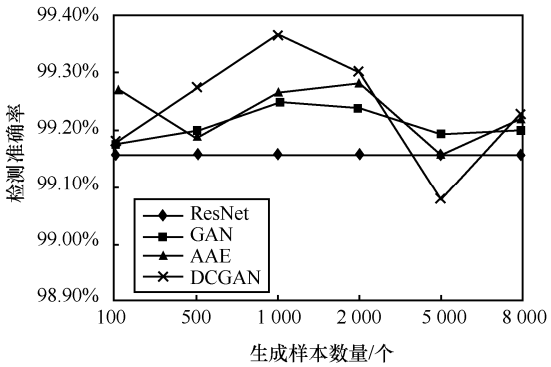


图 5 不同检测模型的检测准确率

如图 5 所示，在混入 100、500、1 000、2 000、5 000、8 000 个生成器生成的样本后，3 种检测模

型的准确率与基于 ResNet 的检测模型相比，均整体小幅提高。从效果来看，基于 DCGAN 的检测模型优于基于 AAE 的检测模型和基于 GAN 的检测模型，特别地，在混入 1 000 个样本时，检测准确率达到最大值，之后随着样本数量增多，准确率开始下降，其原因在于网络流量类型的多样性导致数据分布不均匀，而且会有很多未知的噪声，以此为基础的 GAN 生成的图片分布也会偏离真实的数据分布。在少量样本添加的情况下，可以一定程度上增强小样本分类数据规模，从而增加准确率。随着样本添加数量增加，噪声分布增多，可能会导致训练时噪声较多的样本数超过原样本数，从而学习到了错误的特征。

基于 DCGAN 的僵尸网络检测模型和基于 ResNet 的检测模型的其他性能指标对比如表 3 所示。

表 3　检测性能指标对比（DCGAN 与 ResNet）

模型	准确率	精度	召回率	F1 分数
DCGAN	0.992 3	0.994 6	0.992 9	0.993 7
ResNet	0.991 6	0.992 8	0.993 7	0.993 2

另外，本文比较了基于 DCGAN 的僵尸网络检测模型和基于 ResNet 的检测模型在良性样本和僵尸网络样本上的准确率，如表 4 所示。

表 4 不同样本检测准确率对比（DCGAN 与 ResNet）

模型	良性样本	僵尸网络样本
DCGAN	0.992 9	0.997 1
ResNet	0.993 7	0.996 1

相较于基于 ResNet 的检测模型，基于 DCGAN 的僵尸网络检测模型准确率提升 0.07%，精度提升 0.18%，召回率下降 0.08%，F1 分数提高 0.05%，对良性样本的检测准确率下降 0.08%，对僵尸网络样本的检测准确率上升 0.1%。总体来说，加入生成对抗网络之后检测准确率和精度有所提升，检测性能有了部分提高，有助于僵尸网络流量的识别。

在全为已知僵尸网络的情况下对基于 BiLSTM-GAN 的检测模型分别混入 100、500、1 000、2 000、5 000、8 000 个生成器生成的样本，观测检测模型的检测性能指标。值得注意的是，由于基本的 GAN 不具备学习时间序列的能力，因此本实验未与基本 GAN 进行检测与对比。基于 BiLSTM-GAN 的僵尸网络检测模型和 3.2 节中所述的基于 BiLSTM 的僵尸网络检测模型的检测准确率如图 6 所示。

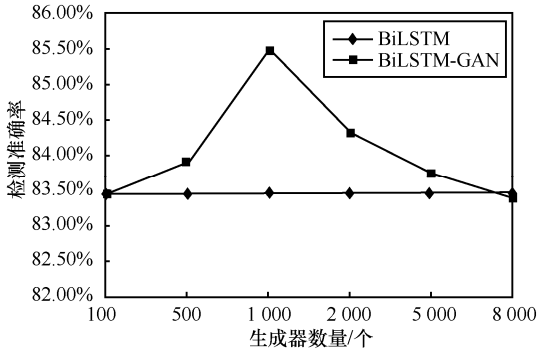


图 6　BiLSTM 与 BiLSTM-GAN 的检测准确率

如图 6 所示，基于 BiLSTM-GAN 的僵尸网络检测模型在混入 1 000 个样本时准确率达到最大，为 85.51%，之后随着样本数量增多，准确率开始下降。混入 1 000 个样本的情况下，基于 BiLSTM-GAN 的僵尸网络检测模型和原有的基于 BiLSTM 的检测模型的其他性能指标对比如表 5 所示。

表 5　检测性能指标对比（BiLSTM 与 BiLSTM-GAN）

模型	准确率	精度	召回率	F1 分数
BiLSTM-GAN	0.855 1	0.915 5	0.823 8	0.867 2
BiLSTM	0.834 5	0.952 7	0.757 7	0.844 1

基于 BiLSTM-GAN 的僵尸网络检测模型和原有的基于 BiLSTM 的检测模型在良性样本和僵尸网络样本上的准确率对比如表 6 所示。

表 6　不同样本的检测准确率对比（BiLSTM 与 BiLSTM-GAN）

模型	良性样本	僵尸网络样本
BiLSTM-GAN	0.823 8	0.959 3
BiLSTM	0.757 7	0.979 4

相较于基于 BiLSTM 的检测模型，基于 BiLSTM-GAN 的僵尸网络模型检测准确率提升 2.47%，精度下降 4.32%，召回率提升 8.72%，F1 分数提升 2.74%，对良性样本检测的准确率上升 6.61%，对僵尸网络样本检测的准确率上升 2.01%。同样地，在加入 BiLSTM-GAN 之后检测性能有一定提高，能够更有效地检测僵尸网络流量。

为研究基于 BiLSTM-GAN 的僵尸网络检测模型中不同类型特征对于检测能力的影响，本文进行了消融实验，按基本特征、异常行为特征和流相似

特征 3 个分类进行研究，如表 7 所示。

表 7 BiLSTM-GAN 模型消融实验

特征	准确率	精度	召回率	F1 分数
基本特征	0.747 3	0.734 1	0.694 7	0.704 0
异常行为特征	0.743 9	0.737 6	0.676 9	0.686 7
流相似特征	0.854 4	0.840 2	0.858 3	0.846 4
基本特征&异常行为特征	0.741 1	0.726 5	0.686 0	0.695 0
基本特征&流相似特征	0.862 3	0.847 7	0.857 5	0.852 0
异常行为特征&流相似特征	0.852 0	0.837 0	0.849 7	0.842 1
全部特征	0.855 1	0.915 5	0.823 8	0.867 2

通过消融实验可以发现，僵尸网络流量中流相似特征对检测能力的影响最大。

另外，实验测试集中僵尸网络样本种类多于测试集，实验统计了基于 DCGAN 的僵尸网络检测模型和基于 BiLSTM-GAN 的僵尸网络检测模型对未知僵尸网络的检测能力。从表 8 中可以看出，本文提出的检测模型对于各种未知僵尸均拥有一定的检测能力。其中，Sogou 和 Smoke 僵尸样本数量本身较少（均为几十个），虽然没有完全检测，但也具备一定的效果。Weasel 僵尸和 Zero Access 僵尸样本数量较多。其中，Weasel 僵尸有上万个，Zero Access 僵尸有上千个，但检测效果较差，原因在于 Weasel 僵尸网络采用 RSA 通信加密，Zero Access 僵尸网络采用 XOR 通信加密，而不是更常见的 RC4 加密方式。Weasel 僵尸网络流量统计特征较特殊，例如其数据流长度均为 900 左右，远高于常见的僵尸网络流量长度。

表 8 未知僵尸网络检测能力

类型	测试集数量/个	DCGAN 检测出的样本数量/个	BiLSTM-GAN 检测出的样本数量/个
Menti	1 300	410	398
Sogou	40	25	20
Mrulo	2 222	725	224
Tbot	593	160	134
Weasel	19 387	89	70
Smoke	36	11	12
Zero Access	463	32	27

5 结束语

本文通过对僵尸网络特性的研究和僵尸网络流量的分析，从空间和时间 2 个维度分别研究僵尸网络特征的生成对抗，实现了基于生成对抗网络的僵尸网络检测算法，提升了检测性能与泛化能力，降低了检测的误报率，具有一定的实用意义与价值。

由于深度学习本身的一些局限性和实际问题的复杂性，深度学习在僵尸网络检测领域的实际应用方面，还存在一些问题。未来的研究工作如下。

基于深度学习的僵尸网络检测模型由于其使用的神经网络层次更深，在提高检测准确率的同时增加了训练的开销，采用的 BiLSTM 模型参数达到百万，ResNet 模型由于层次更深，参数高达千万，对于检测系统的硬件有较高的要求。同时，增加了训练模型的时间，因此在部署到高速或大容量的网络中时，通常无法对网络流量做出全面分析，对于僵尸网络检测的实时性和准确性造成了不利的影响。未来工作将重点研究采用分布式架构或采用分批次进行模型训练，以提高其检测效率。

基于深度学习的僵尸网络检测模型多是在僵尸网络发动攻击后，通过网络流量进行分析检测，不具备很好的僵尸网络攻击发现能力。因此，未来工作将研究在僵尸网络发起恶意活动之前进行检测，如在僵尸网络传播阶段，通过强化学习等算法进行检测，以进一步提高网络安全。

基于统计特征的僵尸网络检测算法由于其统计特性需要人工筛选，因此一定程度上增加了算法的复杂性，且人工筛选的特征在全面性和有效性上仍存在一些问题，而且容易受到恶意对抗攻击的影响。因此，未来的工作是在更加全面地研究统计特征的基础上，研究更稳健的僵尸网络特征和特征提取方法。

基于生成对抗网络的僵尸网络检测模型只采用了传统的训练参数，未根据僵尸网络流量特征的特性设定特定的训练参数。因此，未来的工作将研究不同的超参数及组合对检测模型性能的影响，另外，还将研究可以更好地学习僵尸网络流量特征的对抗生成网络模型对检测效果的影响。

参考文献：

[1] CenturyLink. 2019 threat report[R]. CenturyLink Black Lotus Labs, 2019.

[2] NAIR H S, VINODH E S E. A study on botnet detection techniques[J]. International Journal of Scientific and Research Publications, 2012, 2(4): 2-4.

[3] ANTONAKAKIS M, APRIL T, BAILEY M, et al. Understanding the

Mirai botnet[C]//26th USENIX Security Symposium. Berkeley: USENIX Association, 2017: 1093-1110.

[4] KESSEM L. The Necursbotnet: a pandora's box of malicious spam[R]. Security Intelligence, 2017.

[5] CHECKPOINT R T. JAFF——a new ransomware is in town, and it's widely spread by the infamous Necursbotnet[R]. Checkpoint Research Team, 2017.

[6] KARL S. Crypto-jacking: how cyber-criminals are exploiting the crypto-currency boom[J]. Computer Fraud & Security, 2018(9): 12-14.

[7] SophosLabs Research Team. Emotet exposed: looking inside highly destructive malware[J]. Network Security, 2019(6): 6-11.

[8] Distil Networks. 2019 bad bot report[R]. Distil Networks, 2019.

[9] WAJEEHA A. Why botnets persist: designing effective technical and policy interventions[J]. MIT Internet Policy Research Initiative, 2019(2): 1-52.

[10] BEEK C, DUNTON T, FOKKER J, et al. Mcafee labs threats report[R]. McAfee Report, 2019.

[11] ESMAEILI S, SHAHRIARI H R. PodBot: a new botnet detection method by host and network-based analysis[C]//2019 27th Iranian Conference on Electrical Engineering. Piscataway: IEEE Press, 2019: 1900-1904.

[12] TOKHTABAYEV A G, SKORMIN V A. Non-stationary Markov models and anomaly propagation analysis in IDS[C]//Third International Symposium on Information Assurance and Security. Piscataway: IEEE Press, 2007: 203-208.

[13] SHARAFALDIN I, GHARIB A, LASHKARI A H, et al. BotViz: a memory forensic-based botnet detection and visualization approach[C]//2017 International Carnahan Conference on Security Technology. Piscataway: IEEE Press, 2017: 1-8.

[14] CREECH G, HU J K. A semantic approach to host-based intrusion detection systems using contiguous and discontiguous system call patterns[J]. IEEE Transactions on Computers, 2014, 63(4): 807-819.

[15] BARUAH S. Botnet detection: analysis of various techniques[J]. International Journal of Computational Intelligence & IoT, 2019, 2(2): 1-7.

[16] GU G F. Botnet detection in enterprise networks[M]. Berlin: Springer, 2011.

[17] YAHYAZADEH M, ABADI M. BotCatch: botnet detection based on coordinated group activities of compromised hosts[C]//7th International Symposium on Telecommunications. Piscataway: IEEE Press, 2014: 941-945.

[18] GU G, PORRAS P A, YEGNESWARAN V, et al. Bothunter: detecting malware infection through ids-driven dialog correlation[C]//USENIX Security Symposium. Berkeley: USENIX Association, 2007: 1-16.

[19] GU G, ZHANG J, LEE W. BotSniffer: detecting botnet command and control channels in network traffic[C]//The Network and Distributed System Security Symposium. Saarland: DBLP, 2008: 1-19.

[20] GU G, PERDISCI R, ZHANG J, et al. Botminer: clustering analysis of network traffic for protocol-and structure-independent botnet detection[C]//Proceedings of the 17th USENIX Security Symposium. Berkeley: USENIX Association, 2008: 1-16.

[21] ZHAO D, TRAORE I, SAYED B, et al. Botnet detection based on traffic behavior analysis and flow intervals[J]. Computers & Security, 2013, 39: 2-16.

[22] KARIM A, SALLEH R B, SHIRAZ M, et al. Botnet detection techniques: review, future trends, and issues[J]. Journal of Zhejiang University SCIENCE C, 2014, 15(11): 943-983.

[23] TORRES P, CATANIA C, GARCIA S, et al. An analysis of recurrent neural networks for botnet detection behavior[C]//2016 IEEE Biennial Congress of Argentina. Piscataway: IEEE Press, 2016: 1-6.

[24] HOMAYOUN S, AHMADZADEH M, HASHEMI S, et al. BotShark: a deep learning approach for botnet traffic detection[M]. Berlin: Springer, 2018.

[25] VINAYAKUMAR R, SOMAN K P, POORNACHANDRAN P, et al. DBD: deep learning DGA-based botnet detection[M]. Berlin: Springer, 2019.

[26] MCDERMOTT C D, MAJDANI F, PETROVSKI A V. Botnet detection in the Internet of things using deep learning approaches[C]//2018 International Joint Conference on Neural Networks. Piscataway: IEEE Press, 2018: 1-8.

[27] MEIDAN Y, BOHADANA M, MATHOV Y, et al. N-BaIoT——network-based detection of IoT botnet attacks using deep autoencoders[J]. IEEE Pervasive Computing, 2018, 17(3): 12-22.

[28] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 770-778.

[29] KIM J Y, BU S J, CHO S B. Malware detection using deep transferred generative adversarial networks[C]//International Conference on Neural Information Processing. Berlin: Springer, 2017: 556-564.

[30] 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, 460/461: 83-102.

[31] YIN C L, ZHU Y F, LIU S L, et al. An enhancing framework for botnet detection using generative adversarial networks[C]//2018 International Conference on Artificial Intelligence and Big Data. Piscataway: IEEE Press, 2018: 228-234.

[32] ZHU F, YE F, FU Y C, et al. Electrocardiogram generation with a bidirectional LSTM-CNN generative adversarial network[J]. Scientific Reports, 2019, 9: 6734.

[33] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv Preprint, arXiv: 1511.06434, 2015.

[34] OORD A, DIELEMAN S, ZEN H, et al. WaveNet: a generative model for raw audio[J]. arXiv Preprint, arXiv: 1609.03499, 2016.

[35] MEHRI S, KUMAR K, GULRAJANI I, et al. SampleRNN: an unconditional end-to-end neural audio generation model[J]. arXiv Preprint, arXiv: 1612.07837, 2016.

[36] MOGREN O. C-RNN-GAN: continuous recurrent neural networks with adversarial training[J]. arXiv Preprint, arXiv: 1611.09904, 2016.

[37] YU Y, SRIVASTAVA A, CANALES S. Conditional LSTM-GAN for melody generation from lyrics[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2021, 17(1): 1-20.

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

[39] 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.

[40] 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.

[41] MIRSKY Y, DOITSHMAN T, ELOVICI Y, et al. Kitsune: an ensemble of autoencoders for online network intrusion detection[J]. arXiv Preprint, arXiv:1802.09089, 2018.

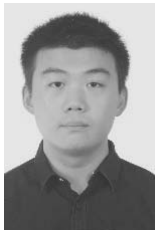
[42] BIGLAR BEIGI E, HADIAN JAZI H, STAKHANOVA N, et al. Towards effective feature selection in machine learning-based botnet detection approaches[C]//2014 IEEE Conference on Communications and Network Security. Piscataway: IEEE Press, 2014: 247-255.

[43] AVIV A J, HAEBERLEN A. Challenges in experimenting with botnet detection systems[C]//4th USENIX Workshop on Cyber Security Experimentation and Test. Berkeley: USENIX Association, 2011: 1-8.

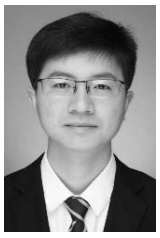
[作者简介]

邹福泰（1973− ），男，江西安福人，博士，上海交通大学高级工程师，主要研究方向为网络威胁感知和网络攻防技术。

谭越（1995− ），男，陕西西安人，上海交通大学硕士生，主要研究方向为网络攻防技术。



王林（1996− ），男，山东济南人，上海交通大学硕士生，主要研究方向为机器学习和威胁情报挖掘。



蒋永康（1996− ），男，贵州遵义人，上海交通大学博士生，主要研究方向为机器学习和恶意软件分析。