深度学习赋能的恶意代码攻防研究进展

冀甜甜[1] 方滨兴[1),2)] 崔翔[2)*] 王忠儒[1)3)*] 甘蕊灵[1)] 韩宇[1)] 余伟强[4)]

[1)](北京邮电大学网络空间安全学院, 北京 100876)

[2)](广州大学网络空间先进技术研究院, 广州 510006)

[3)](中国网络空间研究院, 北京 100010)

[4)](北京丁牛科技有限公司, 北京 100081)

摘 要 深度学习赋能的恶意代码攻防研究已经成为网络安全领域中的热点问题。当前还没有针对这一热点问题的相关综述，为了及时跟进该领域的最新研究成果，本文首先分析并总结了恶意代码攻击的一般流程。对应不同的阶段，本文对深度学习的赋能攻击点和赋能防御点进行了定位，将深度学习助力攻击的技术分为 5 类：（1）基于对抗样本生成的自动化免杀、（2）基于自然语言生成的自动化网络钓鱼、（3）基于神经网络的精准定位与打击、（4）基于生成对抗网络的流量模仿、（5）基于黑盒模型的攻击意图隐藏，并将深度学习助力防御的新型技术分为 3 类：（1）基于深度学习的恶意代码查杀、（2）自动化网络钓鱼识别、（3）深度学习赋能的恶意行为检测；其次，基于以上分类，本文对恶意代码攻防研究中的前沿技术进行了综述，并从技术原理、实际可行性、发展趋势等不同的角度对这些技术进行了深入剖析；再者，由于深度学习的伴生安全问题与其在恶意代码攻防领域的赋能安全问题紧密相关，本文对其中代表性的模型后门攻击与防御的相关技术也进行了关注；之后，本文分析并总结了当前深度学习赋能的恶意代码攻防研究领域中的主要研究方向，并对其未来的发展趋势进行了讨论；最后，深度学习赋能恶意代码攻防研究才刚刚起步，基于恶意代码攻击链的更多可能的赋能攻击与防御点有待研究者继续探索和发掘。

关键词 恶意代码；深度学习；赋能攻击；赋能防御；攻击链

Research on Deep Learning-Powered Malware Attack and Defense Techniques

JI Tian-Tian[1)] FANG Bin-Xing[1),2)] CUI Xiang[2)*] WANG Zhong-Ru[1),3)*] GAN Rui-Ling[1)] HAN Yu[1)] YU Wei-Qiang[4)]

[1)](Key Laboratory of Trustworthy Distributed Computing and Service (BUPT), Ministry of Education, Beijing University of Posts and Telecommunications, Beijing 100876)

[2)](Cyberspace Institute Advanced Technology, Guangzhou University, Guangzhou 510006)

[3)](Chinese Academy of Cyberspace Studies, Beijing 100010)

[4)](Beijing DigApis Technology Co., Ltd, Beijing 100081)

Abstract The research on deep learning-powered malware attack and defense techniques has become a hot issue in the field of cybersecurity. At present, there is no relevant review on this burning issue. In order to follow up on the latest research results in this field, this article first analyzes and summarizes the general malware attack process. Corresponding to different stages, this article locates the attack points and defense points powered by deep learning. The deep learning-assisted attack technologies are divided into five categories: (1) Automated

本课题得到广东省自然科学基金(No. 2019B010136003)、广东省重点研发计划资助项目(No. 2019B010137004)、北京邮电大学博士生创新基金资助项目(No. CX2019115)资助. 冀甜甜，女，1995年生，博士，主要研究领域为网络安全、信息安全.E-mail: jitiantian0728@gmail.com..方滨兴，男，1960年生，博士，中国工程院院士，主要研究领域为计算机体系结构、计算机网络、信息安全.E-mail: fangbx@bupt.edu.cn.崔翔（**通信作者**），男，1978年生，博士，教授，主要研究领域为网络安全、信息安全.E-mail: cuixiang@gzhu.edu.cn. 王忠儒（**通信作者**），男，1986年生，博士，高级工程师，主要研究领域为人工智能、网络安全.E-mail: wangzhongru@bupt.edu.cn.甘蕊灵，女，1996年生，硕士，主要研究领域为网络安全、信息安全.E-mail: lynngan39@163.com.韩宇，女，1995年生，硕士,主要研究领域为网络安全、信息安全.E-mail: hany@bupt.edu.cn. 余伟强，男，1979年生，硕士，主要研究领域为网络安全、信息安全、人工智能.E-mail: yuweiqiang@digapis.cn.

virus evasion based on adversarial sample generation, (2) Automated phishing based on natural language generation, (3) Pinpoint and strike based on neural networks, (4) Traffic imitation based on generative adversarial networks, (5) Hide attack intent based on black-box model; and the deep learning-assisted defense technologies are divided into three categories: (1) Anti-malware based on deep learning, (2) Automated phishing recognition, and (3) Malicious behavior detection powered by deep learning. Secondly, based on the above classification, this article reviews the cutting-edge technologies in this field. Also, it analyzes these technologies in depth from different perspectives, such as technical principles, practical feasibility, and development trends. Furthermore, due to the accompanying security issues of deep learning is closely related to deep learning-powered malware security issues, this paper also pays attention to the accompanying security issues of deep learning and discusses the representative backdoor attack and defense technologies in this field. After that, this article analyzes and summarizes the main research directions in the field of deep learning-powered malware attacks and defenses, and discusses its future development trend. Finally, the research on deep learning-powered malware attacks and defenses is in its infancy, and more possible powered attack and defense points based on malware attack chains remain to be explored by researchers.

Key words malware; deep learning; AI-Powered Attack; AI-Powered Defense; Attack Chain

1 引言

“恶意代码（Malware）”主要包括病毒（Virus）、蠕虫（Worm）、远控木马（Remote Access Trojan，RAT）、僵尸程序（Bot）、勒索软件（Ransomeware）等攻击形态。自 1988 年 Morris 蠕虫出现以来，恶意代码的破坏力开始引发国际关注。从近年来网络安全厂商和媒体报道的重大安全事件报告中可以发现，大多数报告实际上是以恶意代码分析为重心的。毫无疑问，恶意代码在网络空间安全中占据重要位置。

从实际案例来看，近年来爆发的僵尸网络（Botnet）、高级持续性威胁（Advanced Persistent Threat，简称 APT）和勒索软件（Ransomware）等重大网络安全事件，大多数是以恶意代码为核心攻击组件并由此造成实质危害。例如，2001 年爆发的 Code Red 蠕虫在不到一周的时间内感染了近 40 万台服务器，造成全球经济损失约 26 亿美元；2006 年出现的 Zeus 僵尸网络至今依然活跃，据统计该僵尸网络拥有高达 100 万台计算机设备，造成了超过一亿美元的损失；2010 年发起的攻击工控系统而且是核设施的 APT 攻击，其核心攻击组件是 Stuxnet 蠕虫，该攻击对伊朗纳坦兹核电站的上千台铀浓缩离心机造成了实质破坏；2015 年造成乌克兰电网断电事故的 APT 攻击，其核心攻击组件是 BlackEnergy 恶意工具包，这是首次导致停电的网络攻击；2017 年 WannaCry 勒索软件全球大爆发，造成损失达 80 亿美元。诸如此类，不胜枚举。

恶意代码对抗一直是国际网络安全厂商关注重点，国内外安全厂商已研发出较为成熟的终端查杀和网络检测系统，如反病毒软件（Anti-virus Software）、主机入侵防护系统（HIPS）、入侵检测与防护系统（IDPS）。然而，近年来人工智能热潮的再度兴起，对恶意代码的免杀、传播、驻留/持久化、隐蔽通信和精准打击等能力产生了显著的助力效应，并由此引发了新一轮的由人工智能赋能的恶意代码攻防研究，这对传统的恶意代码攻防技术发展将产生重要影响。

人工智能技术对恶意代码发展的影响可分为两种，分别是“赋能效应”和“伴生效应”。所谓赋能效应，主要体现在两个方面，一是指人工智能技术很强大，可以助力恶意代码研发和利用，引发更大的危害，笔者将其称为“赋能攻击”（AI-Powered Attack）；二是指人工智能技术也可以助力恶意代码防御，让安全问题借助人工智能技术得到更好的解决，笔者将其称为“赋能防御”（AI-Powered Defense）。所谓伴生效应，是指尽管人工智能技术在酝酿之初会根据以往的经验去充分地考虑安全问题，但在推出之后，势必会在不断的应用中发现新的脆弱性伴生而来。

本文重点关注赋能效应，并聚焦于人工智能领域中一个重要分支—深度学习技术。在下文中，本文将围绕恶意代码，对深度学习在“助力恶意代码攻击”和“助力恶意代码防御”两个方面的最新研究工作展开综述和分析。值得一提地是，恶意代码的存在形式有多种，涵盖二进制、JavaScript、PowerShell 等，但当前的很多研究[1-3]表明，这些类

型的恶意代码均可通过深度学习进行处理，因此它们均被包含在本文的研究范畴之内，这也进一步突显了本文综述的一般性和通用性。

助力攻击方面：通过对恶意代码攻击链的分析，能够更加清晰地定位可赋能的攻击阶段。针对不同攻击阶段的行为特点，可以发现面向不同攻击阶段的赋能技术，有效地实现对这些技术的分类。安全研究人员已经提出了多种知名的网络威胁框架，其中，Cyber Kill Chain①、MITRE ATT&CK②和NSA/CSS 网络威胁框架 v2（NSA / CSS Cyber Threat Framework v2，简称 NTCTF v2）③三者具有很高的权威性。本文基于以上网络威胁框架和知名网络安全事件，总结形成恶意代码攻击的一般流程（以下称为“恶意代码攻击链”或简称“攻击链”，Attack Chain），如图 1 所示。攻击链包括七个阶段，分别为：准备（Preparation）、投递（Delivery）、突破（Engagement）、存在/持久化（Presence/Persistense）、影响（Effect）、命令与控制（Command and Control，简称 C2）和规避（Evasion）。与之对应地，本文将深度学习助力攻击的赋能技术分为九类，分别是：基于对抗样本生成（Adversarial Sample Generation）的自动化免杀、基于自然语言生成（Natural Language Generation，简称 NLG）的自动化网络钓鱼、基于深度学习分类的精准定位与打击、基于生成对抗网络（Generative Adversarial Network，简称 GAN）的流量模仿、基于黑盒模型的攻击意图隐藏、自动化漏洞挖掘、自动化漏洞利用、自动化绕过凭证和基于深度学习的密码破解。其中自动化漏洞挖掘与利用，以及自动化绕过凭证和密码破解技术分属于一个独立的研究领域，且已有大量的工作和研究进展的跟踪，故本文不将它们作为关注的重点，而是重点关注与恶意代码紧密相关其他 5 类研究工作。

① Cyber Kill Chain, https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html

② MITRE ATT&CK, https://attack.mitre.org/

③ NSA / CSS Cyber Threat Framework v2, https://media.defense.gov/2019/Jul/16/2002158108/-1/-1/0/CTR_NSA-CSS-TECHNICAL-CYBER-THREAT-FRAMEWORK_V2.PDF 2018,11,29

恶意代码攻击的一般流程（恶意代码攻击链）

阶段	准备 Preparation	投递 Delivery	突破 Engagement	存在/持久化 Presence/Persistense	影响 Effect	命令与控制 Command and Control	规避 Evasion
Cyber Kill Chain	√	√	√	×	√	√	×
MITRE ATT&CK	×	×	√	√	√	√	√
NSA/CSS NTCTF v2	√	√	√	√	√	√	√
深度学习赋能攻击技术	①⑥	②	⑦	⑧⑨	③	④	④⑤
深度学习赋能防御技术	—	b	ad	ad	c	c	—

注，序号分别表示如下：
①：基于对抗样本生成的自动化免杀，②：基于自然语言生成的自动化网络钓鱼，
③：基于深度学习分类的精准定位与打击，④：基于生成对抗网络的流量模仿，
⑤：基于黑盒模型的攻击意图隐藏，⑥：自动化漏洞挖掘，
⑦：自动化漏洞利用，⑧：自动化绕过凭证，
⑨：基于深度学习的密码破解
a：基于深度学习的恶意代码查杀，b：自动化网络钓鱼攻击识别，
c：深度学习赋能的恶意行为检测，d：自动化漏洞修复

图 1 基于恶意代码攻击链的赋能技术分类

助力防御方面：假设深度学习赋能的恶意代码攻击实际可行，安全防御人员在理论层面也进行了很多超前探索，因此本文对当前最新的深度学习赋能的恶意代码防御研究工作进行了总结归纳，以期推动赋能恶意代码防御的研究进展。同理，由图 1 所示，我们不关注与自动化漏洞挖掘与利用对应的自动化漏洞修复技术，而是重点关注基于深度学习的恶意代码查杀、自动化网络钓鱼攻击识别、深度学习赋能的恶意行为检测三类赋能防御技术，旨在从防御者可检测的目标侧对恶意代码攻击链的各个环节开展防御研究。

综上，将本文的贡献总结如下：

（1）通过对网络威胁框架和知名网络安全事件的分析，本文提取并总结了具有通用性的恶意代码攻击链，通过刻画攻击流程，帮助有效定位恶意代码攻防赋能点；

（2）以攻击链为主线，本文关注深度学习在恶意代码攻防研究中的赋能效应，对深度学习助力恶意代码攻击和防御两个方面的研究工作进行了分析、归纳、总结与思考；

（3）本文基于深度学习的助力安全问题进行了扩展延伸，总结并分析了模型后门攻击与防御的相关研究工作，它属于深度学习的伴生效应，但在恶意代码“投递”阶段也起到助力安全的作用；

（4）本文对恶意代码攻防研究的未来发展趋势进行了分析与展望，旨在从更前沿的视角研究基于恶意代码的智能安全问题。

具体地，本文的组织结构如下：第 2、3 章分别对深度学习赋能的恶意代码攻击和防御的相关研究进行了梳理和总结；第 4 章扩展综述了深度学习的伴生安全研究；第 5 章概括总结了全文工作，并对深度学习赋能恶意代码攻防研究的未来发展趋势进行了分析与讨论。

2 深度学习助力攻击

在深度学习助力攻击方面，攻击者试图基于恶意代码攻击链对可操作的各个攻击环节进行赋能，以增强攻击的鲁棒性。基于恶意代码攻击链刻画的攻击流程，本文将深度学习赋能研究的新型攻击技术具化为 5 类，它们分别是：①基于对抗样本生成的自动化免杀、②基于自然语言生成的自动化网络钓鱼、③基于深度学习分类的精准定位与打击、④基于生成对抗网络的流量模仿、⑤基于黑盒模型的攻击意图隐藏。

2.1 基于对抗样本生成的自动化免杀

在恶意代码攻击过程的“准备”阶段中必要的一件事情是恶意代码构建，以提升恶意代码的免杀和生存能力。而每当恶意代码出现新的趋势和威胁时，反病毒引擎作为与恶意代码对抗过程中的产物则需要不断发展以制衡恶意代码。当前反病毒引擎在变种检测方面的健壮性正在不断加强，基于深度学习的反病毒引擎已经面世，例如，腾讯开发的 TRP-AI 反病毒引擎就是基于深度学习来查杀恶意

代码的。但相应地，反病毒引擎的发展也必定会促进恶意代码的研究。为了深入了解攻击者的意图，并提前做好防御措施，安全研究人员尝试以攻击者视角预测未来恶意代码的发展趋势，以期提前洞悉恶意代码的范式转变。

在恶意代码免杀方面，深度学习对比机器学习的优势在于：1）深度学习会自动学习到重要的特征，不需要人工选择特征；2）对于恶意代码中某些潜在且关键特征的增加或删除，深度学习可以通过自主学习实现自适应，在恶意代码查杀的可扩展性上也优于机器学习的方法。因此，从开发新型恶意代码的角度出发，一种深度学习赋能免杀的对抗性恶意代码被提出，用于实现基于对抗样本生成的自动化免杀。

2016 年，Kathrin Grosse[4]等人基于前向导数算法生成具有对抗性的恶意代码示例，以对抗深度神经网络（Deep Neural Network，简称 DNN）的查杀，通过实验证明，该方法对恶意代码实现了 85%的误分类率，验证了基于对抗样本生成的恶意代码攻击的可行性。2017 年，Weiwei Hu[5]基于 GAN 提出了 MalGAN 模型来生成对抗性恶意代码，以绕过黑盒检测系统，实验结果表明 MalGAN 能够将检测率降低到接近零，并使基于再训练的防御性方法难以对抗此类攻击；同年，Weiwei Hu[6]等人还将深度学习中的递归神经网络（Recurrent Neural Network，简称 RNN）与 GAN 相结合，在原始恶意代码的 API 序列中插入一些不相关的 API，生成基于顺序的对抗性恶意代码，可有效对抗多种不同 RNN 结构的的模拟反病毒引擎。2018 年，Bojan Kolosnjaji[7]等人首次在字节粒度上提出在恶意代码末尾填充字节的方式来生成对抗性恶意代码，并基于梯度下降算法指导决定要填充的字节，其测试准确率高达 92.83%，并在实际对抗基于字节粒度的检测系统 MalConv[8]时获得了 60%的成功率；同年，Hyrum S. Anderson[9]等人利用深度强化学习网络（Deep Reinforcement-learning Network，简称 DRN），提出了一种基于对抗样本生成的黑盒攻击方法，用于攻击静态 PE 反杀毒引擎，这是当前第一个可以产生对抗性 PE 恶意代码的工作，在模拟现实的攻击中达到了 90%的成功率。

表 1　深度学习赋能的自动化免杀技术对比与分析

年份	文献	前提/假设	验证引擎	数据集（来源）	测试成功率	核心方法	实际可行性
2016	[4]	白盒假设（攻击者知道神经网络模型的结构、参数）	DNN	DREBIN Android	85%	基于攻击神经网络的前向导数算法	否
2017	[5]	黑盒假设，但攻击者知道恶意代码检测算法使用的功能	GAN	https://malwr.com/	接近 100%	替代检测器拟合黑盒检测系统	否
2017	[6]	黑盒假设（攻击者不知道神经网络模型的结构、参数）	RNN、GAN	https://malwr.com/	96.97% – 99.56%	不相关的 API 序列插入	否
2018	[7]	黑盒假设	DNN	VirusShare, Citadel, APT1	92.83%	字节粒度的梯度下降算法	否
2018	[9]	黑盒假设	DRN	Virus Share, VirusTotal	90%（AUC: 99.3%）	深度强化学习算法做预测与策略评估。	否

以上基于对抗样本生成的自动化免杀方法中，API 或字节填充等操作是直接反馈到恶意代码的样本中的，体现在深度学习的赋能效应上则是恶意代码的免杀和生存能力的增强，表 1 中的测试成功率便很好地证明了深度学习在该类技术上赋能的有效性。此外，这些自动化免杀方法，它们分别从不同的角度为生成对抗性恶意代码提供了不同的攻击思路。为了深入了解并探索该类新型赋能攻击技术的实际可行性，以及其未来的发展趋势，从不同维度对这些方法做深度分析与对比是必要的。如表 1 所示，本文从这些方法基于的前提/假设、使用的验证引擎、实验数据集、实验结果等几个角度进行了概述。从该表中可以看出基于对抗样本生成的自动化免杀正在不断发展进步，但同时也面临着很多问题，具体总结如下。

（一）三个方面的发展与进步：

1）自动化免杀这一研究的限制条件越来越严苛，研究从最初的白盒假设上升了到黑盒假设，这意味着研究者对于攻击的预测也越来越偏向于实际应用场景，有利于防御者提前构建面向实际应用的新型防御措施，掌握攻防主动性。

2）验证引擎中使用的深度神经网络结构也已

经从单纯的 DNN 涵盖到包含 GAN、RNN、DRN 等多种网络模型。从广义上讲，DNN 泛指包含了全连接、深度卷积神经网络（Convolutional Neural Network，简称 CNN）、RNN 等具体变种形式的神经网络结构，但有时研究者也用于特指全连接的神经网络结构，具体取决于不同文献中对 DNN 定义的不同；CNN 的主要功能在于特征提取；RNN 主要用于序列生成；GAN 的主要作用是序列生成和鉴别；DRN 则用于预测和评估。由表 1 可以看出，以上这些深度学习模型均可用于自动化免杀技术的攻击验证，可见基于对抗样本的自动化免杀已经建立了对模型结构的普适性，并可针对多种模型架构的反病毒引擎实施攻击，使得该类攻击的实际可操作性进一步增强。

3）恶意代码的生成方式除了涵盖粗粒度的功能添加和删除，还扩展到了字节填充等细粒度的方法上，并且通过实验测试，对抗性恶意代码成功绕过验证引擎的概率也越来越高，有的甚至达到了 100%，从技术细节和攻击性能上为防御者提供了启发和警示的作用。

（二）两个方面的问题与挑战：

第一，当前恶意样本多通过爬虫、复现等方式来获取，但由于恶意样本存在获取难、运行难、易失去活性，而且数量少等问题，各项研究中使用的数据集比较单一、分散，还没有一个统一、可公开共享的大规模数据集供研究使用。

第二，这些自动化免杀方法在理论研究中往往使用模拟的反病毒引擎进行测试，其中模拟反病毒引擎的典型生成方式是：攻击者利用可能的测试数据作为输入，从真实的反病毒引擎中获取输出，然后基于已知的输入和输出来训练模拟引擎，以替代实际的反病毒引擎。虽然测试成功率很高，但在实际应用中并不可行，具体有两个原因：1）数据集的限制，使得模拟的反病毒引擎训练不充分，有限的验证也并不足以代表其实际的免杀效果；2）当前对抗性恶意代码样本的生成并不改变原始样本携带的恶意行为，而在真实环境中部署的反病毒引擎有很多检测维度，其中不仅包括了对恶意代码中敏感字符串的检查，还会模拟执行恶意代码进行安全检测，一旦恶意行为暴露，这种用于免杀的恶意代码也会被发现。因此，在面对拥有多维度检测视角的反病毒检测引擎时，对抗性恶意代码相比原始恶意代码只是多绕过了其中一个维度的检测，两者在实际反杀效果上的能力表现上也有待进一步地实验评估与验证。

综上，本文总结得出结论：当前基于对抗样本生成的自动化免杀在实际应用中并不可行。然而，即使如此，作为恶意代码攻击“准备”过程中增强免杀性和生存性的一个先进的技术手段，深度学习赋能的自动化免杀技术仍具有重要的研究价值。一方面，该类赋能攻击手段与恶意代码攻击链中的“准备”阶段相对应，虽然实际并不可行，但该方向在未来将不可避免地成为攻击者在武器构建中继续深耕的领域，安全防御人员站在攻击者的角度对该类攻击技术的不断探索和预测，将有利于更加深入且多方位地了解潜在的攻击意图；另一方面，由于深度学习模型的可迁移性，基于对抗样本生成的自动化免杀的研究成果在一定程度也可被迁移应用，这些具有迁移性的研究成果不仅可以被用于恶意代码过程的其他环节中，也可以启发对安全防御技术的探索。

最后，对此类基于对抗样本生成的自动化免杀技术，其基本防御方法主要在下文中“基于深度学习的恶意代码查杀”和“深度学习伴生的模型后门防御”两类防御技术中讨论，这两类技术也可以结合使用，以从不同的维度增强恶意代码检测模型的鲁棒性。

2.2 基于自然语言生成的自动化网络钓鱼

在“准备”阶段完成武器构建之后，下一步攻击者需要将武器投放至目标环境中，这就是在“投递”阶段需要完成的任务。网络钓鱼是在“投递”阶段中主流的攻击方式。据统计，趋势科技的安全研究人员在 2012 年发现，91％的定向攻击（Targeted Attack）[10]用到了鱼叉式网络钓鱼攻击手法，能够成功诱骗受害者打开恶意文件或网站；网络钓鱼威胁管理公司 PhishMe 在 2017 年报告了相同的数字（91％）[11]，并指出网络钓鱼仍是第一大攻击媒介；之后，Verizon 2019 DBIR 数据泄露报告[12]称，有 32％的数据泄露事件归结为网络钓鱼，78％的网络间谍事件涉及网络钓鱼。另外，本文通过对实际攻击案例的分析发现，APT 是最主要的定向攻击，尤其在大部分 APT 组织采用的恶意代码投递方式中，鱼叉式网络钓鱼是他们的首选技术，例如，2013 年，一封假冒银行交易的网络钓鱼信件导致韩国爆发史上最大的 APT 攻击[13]。

当前来自网络钓鱼攻击的威胁持续增长，但传统的网络钓鱼已经被证明可以实现成功对抗，因此已经有很多攻击者和安全研究人员开始关注更高

级的自动化网络钓鱼方法。一方面，网络钓鱼特别适合 NLG 方法，利用 NLG 方法，可以利用重复出现的文本模式来识别感兴趣的主题并生成目标可能响应的文本内容；另一方面，深度学习为 NLG 的应用带来了诸多有益效果：1）有效降低了输入特征的维度，从而降低了输入层的复杂性；2）具有其他浅层模型不可比拟的灵活性，模型更复杂，能够对数据进行更精准的建模，从而增强实验效果；3）将词语、文本由词空间映射到语义空间，一定程度上缓解了语义鸿沟的问题。因此，在新型自动化网络钓鱼的研究中，研究者更倾向于使用以深度学习模型为框架的 NLG 技术展开研究，其中主要包括以电子邮件和社交网站作为恶意代码传输载体的新型网络钓鱼攻击技术，具体概括如下：

首先，鱼叉式钓鱼软件是 APT 攻击中最有效的手段。2017 年，Shahryar Baki[14]等人设计了一种使用 NLG 的可扩展方案来开展电子邮件伪装的鱼叉式钓鱼攻击，并且他们针对两个知名人物（Hillary Clinton 和 Sarah Palin，分别简称为 HC 和 SP）验证了鱼叉式钓鱼攻击的效果，攻击结果显示对 HC 的电子邮件进行的鱼叉式钓鱼攻击欺骗了 34%的人，对 SP 的攻击欺骗了 71%的人。2019 年，Avisha Das[15]等人使用基于 RNN 的 NLG 技术，通过在训练过程中注入恶意意图（例如欺骗性的网页链接或超链接），并在合成电子邮件中生成恶意内容，从而实现了一个进行高级电子邮件伪装攻击的 RNN 原型系统，该系统的训练依赖于两个数据集：一个是合法电子邮件数据集，主要是从真实个人的发件箱和收件箱中提取的；另一个是网络钓鱼邮件数据集，由个人收集的 197 个网络钓鱼电子邮件，以及 Jose Nazario 网络钓鱼语料库中的 3392 封网络钓鱼电子邮件组成。最终通过评估实验证明，对比 Shahryar Baki 等人使用的 Dada 引擎，RNN 生成的伪电子邮件具有更好的连贯性和更少的语法错误，能更好地作为网络钓鱼电子邮件进行攻击。

其次，以 Twitter 为代表的社交网站可以允许访问大量的个人数据，有利于钓鱼攻击者探索感兴趣的攻击目标；而这些社交网站又因为具有对机器友好的 API，通俗易懂的语法以及普遍存在的缩址链接，因此成为了恶意代码传播的理想场所。具有代表性的一项研究是 John Seymour[16, 17]等人在 2016 年 Black Hat 大会上提出的基于 Twitter 自动化的端到端鱼叉式网络钓鱼，他们实现了第一个可以针对高价值用户生成网络钓鱼帖子的生成器 SNAP_R（Social Network Automated Phishing with Reconnaissance）。SNAP_R 基于长短时记忆（Long Short-Term Memory，简称 LSTM）神经网络模型实现，使用鱼叉式网络钓鱼渗透测试数据进行训练。一方面，为了使点击率更高，SNAP_R 会动态植入从目标用户以及他们转发或关注用户的时间轴帖子中提取的主题；另一方面，SNAP_R 通过聚类来扩展模型，它根据用户的整体数据对用户进行分类，以确定高价值目标用户。最后，实验证明该方法达到了很高的成功率（30%-60%），是传统电子邮件攻击准确性（5%-14%）的三倍，并且一度胜过了人工执行相同任务的人员（准确性约为 45%）。

综上，从上述研究可以看出，在该类技术中，深度学习的赋能效应体现为恶意代码投递能力的提升，尤其是基于 NLG 的自动化网络鱼叉式钓鱼正在不断发展，可实现恶意代码投递的自动化、规模化。本文围绕这些研究，就深度学习的赋能效应进行了深入分析，从中总结得出三个结论：

第一，攻击者越来越倾向于进行自动化的网络钓鱼攻击，以使用低成本来获取高收益。很多理论研究成果已经开始向实际应用做转化，例如 2016 年的 Kiwicon 黑客大会上，意大利安全专家 Michele Orru 发布了一款自动化网络钓鱼工具 PhishLulz[18]，并且在对澳大利亚官员调查测试中，成功欺骗了 40%的澳大利亚公务员；2019 年的 Hack in the Box 安全大会上，安全专家已经证实，通过协同使用 Muraena 和 NecroBrowser[19]，自动化的钓鱼攻击可以穿透双重验证，实现攻击的自动化。

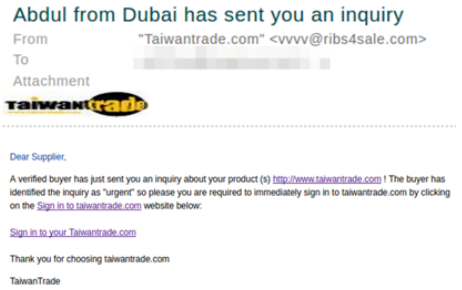

图 2 钓鱼邮件示意图

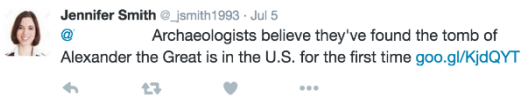

图 3 社交网络钓鱼帖子示意图

第二，如图 2 和图 3 所示，基于深度学习的 NLG 技术可以根据目标用户的邮件或帖子内容的上下文生成自然语言文本，攻击者可以使用此类系统生成涵盖邮件、推文等十多种类型的文本内容，

这类 NLG 系统很有可能成为网络钓鱼攻击者手中的危险工具。因此，这类深度学习赋能攻击技术需要引起安全防御人员的充分重视，而且对网络钓鱼深层且全面的查杀需要面向云端进行，邮件服务提供商和社交网站服务提供商也需要进一步改进他们的产品。

第三，潜在的安全威胁需要引起重视。有研究估计，网络总流量的 61.5%来自机器人；且 Twitter 的一项研究表明，在最活跃的帐户生成的 Twitter 帖子中，机器人占了 32%①。这两项数据表明机器人已经擅长于模仿人类行为来传播信息。然而，如果大量的网络机器人拥有类似 SNAP_R[16]的功能，并以网络钓鱼攻击为目的，就可以针对目标用户生成其感兴趣的钓鱼帖子（比如可以在一定程度上影响或操纵网民心理的舆论信息）来开展攻击。依照目前 SNAP_R 的攻击成功率进行预测，如此大规模的自动化网络钓鱼攻击将会带来严重的安全威胁；而且随着深度学习的不断发展，基于深度学习构建的僵尸机器将难以检测。因此，本文提议进行更多的网络安全性研究，尤其是针对鱼叉式网络钓鱼攻击，需要新型的防御手段来检测隐藏或伪装的钓鱼机器人。

2.3 基于神经网络的精准定位与打击

恶意代码在投放至目标环境，并建立了持久化控制之后，主要目的就是为了在“影响”阶段释放恶意攻击行为。由于这些攻击行为发生在终端受害者主机上，防御者部署检测时更具有主动性，相比之下，该阶段留给攻击者探索赋能攻击的空间并不大。但攻防竞赛总是不断发展制衡的，在针对特定目标实现精准打击的场景中，攻击者发现了一种新型、有力的攻击手段，即将深度学习模型作为生成攻击的一个核心组件之一，利用深度学习的分类功能实现对攻击目标的精准定位与打击。深度学习模型作为攻击组件，与攻击载荷一起嵌入到恶意代码体内，它是恶意代码是否释放恶意行为的决策者，只有在特定攻击目标出现时，深度学习模型才会决定释放恶意行为，实现精准打击。因此，深度学习模型可赋能恶意代码的精准识别能力。

当前这类赋能攻击中最具代表性的研究成果就是 IBM 研究院在 2018 年的 Black Hat 大会上展示的一种 AI 赋能的恶意代码 — DeepLocker[20]，它借助 CNN 模型实现了对特定目标的精准定位与打击。

本文通过实验再现了精准定位与打击这类攻击手段的实际可行性，并对其核心思想进行了深度剖析。具体地，精准定位与打击由精准定位和精准打击两个关键步骤构成：

1）对于精准定位，其在攻击中的主要目的是对特定攻击目标的精准识别，深度学习模型的分类功能便足以完成此项任务。深度神经网络分类器的实现是利用已知的训练数据、给定的类别来学习分类规则，然后对未知数据进行分类或预测。

如图 4 所示，以简单的二分类人脸识别为例，假定“周星驰”的面部图像为识别目标，建立对应类别为“是周星驰”（机器标签为“1”）和“不是周星驰”（机器标签为“0”）两种面部图像数据集（比如周星驰的面部图像数据集和非周星驰的面部图像数据集）。

根据表 2 的混淆矩阵，一般可以用公式(1)的精准率（Precision，表示为 P）和公式(2)的召回率（Recall，表示为 R）两个指标对已训练好的神经网络模型的预期效果进行评价。在达到预期指标的前提下，如图 4 所示的二分类神经网络便可以很好地实现对未知数据的预测，即对大部分周星驰的面部图像均可以正确分类为“是周星驰”（用概率表示，则 TP 趋近于 1，FN 趋近于 0），而对大部分非周星驰的图像也可以正确分类为“不是周星驰”（用概率表示，则 TN 趋近于 1，FP 趋近于 0），如此便实现了对特定目标（周星驰面部图像）的精准定位。

表 2 基于二分类神经网络的混淆矩阵

分类	真实样本 1（周星驰）	真实样本 0（非周星驰）
预测 1（是周星驰）	TP	FP
预测 0（不是周星驰）	FN	TN

$$P = \frac{TP}{TP+FP} \tag{1}$$

$$R = \frac{TP}{TP+FN} \tag{2}$$

2）精准打击建立在精准定位目标的基础上，本文在实验过程中，首先通过给定特定目标，提取二分类神经网络模型中特定全连接层的输出，该输出被用作对称密钥来加密攻击载荷；然后，本文将二分类模型与加密后的攻击载荷一起，嵌入在某一正常应用程序中，以实现智能恶意代码的构造。

当智能恶意代码部署在受害主机上并启动运

① An In-Depth Look at the Most Active Twitter User Data, https://sysomos.com/inside-twitter/most-active-twitter-user-data/ 2009,8

行时，二分类神经网络模型会同步启动对目标的检测。在非周星驰的面部图像出现的情况下，智能恶意代码未检测到目标，便会如正常应用程序一般发挥发挥其正常功能；只有在是周星驰的面部图像被检测到时，二分类模型才会给出其特定全连接层的正确输出，这时该输出被用作对称密钥来解密攻击载荷，进而攻击载荷释放攻击行为，实现精准打击。

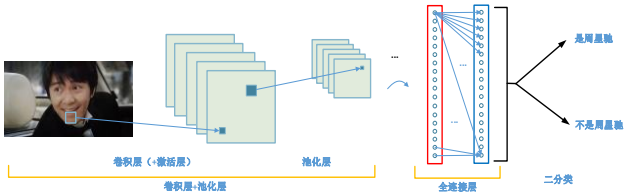

图 4 基于深度神经网络的特定目标识别示意图

通过对 DeepLocker 的模拟实验，本文在深入剖析了精准定位与打击的攻击原理之后，并对其在受害主机侧完成部署后的防御对抗难度进行了分析，总结如下：

第一，在这种新型的攻击中，攻击者使用深度学习模型来识别的目标特征有很多，可以是人脸、声音、用户行为、传感器、地理位置、物理环境、软件环境，以及非常可靠的虹膜等等。站在对抗防御的角度，由于深度学习模型的可迁移性，其应用场景有很多，即使安全防御人员在受害主机上检测到了深度学习模型的存在，也很难知道模型作用于攻击时的应用场景，从场景的遍历上增加了防御的难度。

第二，即使防御者知道了作用于攻击的深度学习模型的应用场景，例如人脸识别，但破解攻击仍是遥不可及的，因为具体攻击目标的面部图像是不可知的，防御者需要想从海量的图像中遍历以发现特定的攻击目标，然而这种海量遍历是一件非常困难且几乎不可能的事情。因此，从精准识别目标的角度，这种新型的攻击一旦被用于实际将难以被防御。

除此之外，本文还基于 ATT&CK 知识库中的线索，对实际攻击案例进行了分析，发现在当前的 APT 攻击组织中，FIN7 所使用的称为 BIOLOAD[①]的恶意软件与 DeepLocker 紧密相关，它是 DeepLocker 思路的一个具体实例，尽管 BIOLOAD 并不是基于深度神经网络模型的实现，但在实际攻击中它却基于 Hash 算法实现了对特定攻击目标的精准定位与打击。具体地，BIOLOAD 是针对其感染的每台计算机量身定制的，如图 5 所示，BIOLOAD 通过获取“COMPUTERNAME”（即计算机名称）实现了对目标计算机的精准定位；并且 BIOLOAD 在密钥获取上并不依赖于远程服务器，而是基于“COMPUTERNAME”来获取用于解密攻击载荷的密钥，进而实现了对特定目标的精准打击。

```
_itow(CRC32(_wdupenv_s(COMPUTERNAME))) + "-" +
_itow(CRC32(_wdupenv_s(PROCESSOR_IDENTIFIER))) + "-" +
_itow(CRC32(_wdupenv_s(USERNAME))) + "-" +
_itow(CRC32(_wdupenv_s(PATHEXT)))
```

图 5 BIOLOAD 密钥连接示意图

综上，当前精准定位与打击的攻击思路已经不仅仅只停留在理论分析层面，在扩展分析中，本文发现该类攻击手法已被攻击者应用于实际的高级持续性威胁攻击中，并且通过对防御难度的分析，本文认为这类攻击技术一旦被攻击者广泛利用，将难以实现对抗。因此该类攻击带来的安全风险需要引起安全研究人员的高度重视，针对这类新型攻击的防御研究当刻不容缓。

2.4 基于生成对抗网络的流量模仿

“影响”阶段中精准定位与打击强调一次攻击的有效性，而在“C2”和“规避”阶段则需要强化攻击的隐蔽性。在“C2”阶段中，攻击者需要保证在攻击过程中对远程控制命令的及时更新，并能维持 C2 信道的隐蔽性。在“规避”阶段中则需要尽可能地增强各种攻击活动的隐蔽性。因此，攻击者希望由 C2 通信和各种恶意攻击活动产生的恶意流量可以通过模仿正常流量实现自适应，以尽可能地避开检测[21]。

在模仿对抗方面，GAN 由 Goodfellow[22]等人在 2014 年提出，到现在已经在各个领域产生了重要影响，尤其是在图像领域，GAN 已经被成功应用于超分辨率图像生成、图像隐写等方面。作为一种新的深度神经网络系统，GAN 生成的数据非常逼真，从某种意义上讲，GAN 为指导恶意代码攻防的研究提供了一种新的思路。因此，从预测攻击的视角出发，为了实现恶意流量的自适应，研究人员提出了基于 GAN 的流量模仿技术。

2018 年，Maria Rigaki[21]等人使用 GAN 来学习模仿 Facebook 聊天流量，通过修改恶意软件的源代码从 GAN 接收参数，通过参数的反馈，可以调整其“C2”信道上基于流量的通信行为，以使其不会被阻止，实验结果表明这种流量模仿的攻击方法可

① Introducing BIOLOAD: FIN7 BOOSTWRITE’s Lost Twin, https://www.fortinet.com/blog/threat-research/bioload-fin7-boostwrite-lost-twin.html 2019,12,26

以成功绕过 Stratosphere IPS。2019 年，Zilong Lin[23]等人提出了一种新的 GAN 框架 IDSGAN，用于在流量维度上针对入侵检测系统进行黑盒对抗性攻击，IDSGAN 由生成器、鉴别器和黑盒 IDS 组成。实验表明，IDSGAN 仅通过修改攻击流量的部分非功能性特征，就可使各种黑盒 IDS 模型的检测率降低到接近 0，提高了入侵的有效性。

此外，由于深度学习应用的可迁移性，本文对另外两类不是以恶意攻击为目的、但同属于流量模仿的研究工作也进行了介绍，以启发攻防对抗的研究思路。

一类关注数据集缺乏问题。2019 年 WooHo Lee[24]等人提出一种基于图像生成攻击流量的方法，该方法专注攻防对抗中数据量非常少的恶意流量，使用 WGAN-GP 模型来增加攻击级别的流量训练数据，通过解决数据集不平衡的问题，用以实现增强的模型学习和分类。同年，Markus Ring[25]等人提出使用 GAN 生成逼真的基于流的网络流量，其中既包含有正常的用户行为流量，也包含着恶意攻击流量，并且实验证明他们提出的方法可以生成高质量的数据。

另一类关注实现以匿名通信为目的的流量伪装，该类流量伪装与实现恶意攻击、逃避检测的目的不同，主要是为了保护通信中的私密信息。2019 年 Jie Li[26]等人提出了一种称为 FlowGAN 的动态流量伪装技术，它包括 GAN 生成器、网络流生成器、本地代理服务器和远程代理服务器。FlowGAN 的核心思想是通过提供源（或审查）流和目标流，FlowGAN 可以自动提取目标流的流量特征，并基于这些特征将源流变形为目标流的形式。基于这一核心思想，匿名通信的流量便可被伪装变形为可规避互联网审查的流量，而且实验结果表明 FlowGAN 在流量伪装方面具有不可区分性和延迟小的特性。

综上所述，基于 GAN 的流量模仿是一个新型的赋能领域，深度学习在该领域中实现的赋能效应是对恶意代码攻击隐蔽性的增强。当前，以攻击者视角在该领域中开展的理论研究仍相对较少，并且该类攻击技术在实际攻击中尚未得到应用，但在基于 GAN 的流量模仿这一攻击理论被提出后，安全研究人员对如何开发新的检测方法尚无明确的想法。因此，本文将流量模仿相关的研究工作进行了分析与归纳，旨在启发安全防御人员，促进安全防御的研究。如表 3 所示，这些工作中既包含了带有攻击企图的流量模仿，也包括了为实现数据集生成和匿名通信而进行的流量模仿，三类研究虽然核心目的不同，但因深度学习的迁移性，这些工作方法都可迁移并应用于攻击场景中。

表 3 流量模仿技术对比与分析

目的	年份	相关文献	前提/假设	模仿引擎	数据集（来源）	约束条件/解决问题
对抗性黑盒攻击	2018	[21]	黑盒假设	GAN、LSTM	Facebook 流量	1）恶意代码是真实的且可以在网络中执行真实的恶意代码操作； 2）对恶意代码的拦截是真实的。
	2019	[23]	黑盒假设	Wasserstein GAN	NSL-KDD	对 IDS 检测系统在抵抗对抗攻击时的健壮性提出质疑
数据集生成	2019	[24]	白盒假设	WGAN-GP	UNSW-NB15 、KDD-NSL	数据不平衡问题，即缺少网络攻击流量
	2019	[25]	白盒假设	WGAN-GP	CIDDS-001	数据集缺乏：1）可用的数据集通常已过时或存在其他缺点；2）对于基于网络的入侵检测，很少公开提供带有标签的数据集。
匿名通信	2019	[26]	黑盒假设	Wasserstein GAN	10000 traffic flows to baidu.com	现有的流量变形/隧道技术表现出强大的流量模式且缺乏动态性，因此很可能被识别。

从表 3 的分析结果，本文总结得出两个结论：

第一，研究者在流量模仿中使用的数据集也是分散、不统一的，这与自动化免杀领域的赋能研究面临着相同的局限问题。即使在该领域已经存在了数据集生成的研究工作，但原始数据集本身不统一、不共享，其模仿生成的数据集也无法实现通用，

这仍然是研究中的一大限制，无论对于攻击的预测和新型防御的提出，都是亟待解决的一个难点。

第二，良性流量模仿将会大大降低当前防御检测的有效性。尽管流量模仿系统要求与恶意代码同在一个受害主机或内网环境之中，对良性流量的模仿一旦被应用于实际的攻防对抗环境中，可导致专注流量检测的网络流量分析（Network Traffic Analysis，简称 NTA）技术完全失效。这是因为采用流量模仿技术生成的恶意流量与正常的良性流量在特征分布、行为表现等方面几乎保持一致，很容易导致流量检测系统误分类，将恶意流量识别为良性流量。站在防御者视角上，如果不能提取出有效且明显区别于良性流量的特征，NTA 技术在流量检测中将面临挑战。

因此，尽管基于 GAN 的流量模仿是一项新型的攻击预测研究，安全防御人员仍需要加快研究步伐，争取在应用流量模仿的实际攻击造成危害之前，能够提前部署可对抗攻击的防御设施。

2.5 基于黑盒模型的攻击意图隐藏

在“C2”和“规避”阶段，攻击者通过流量模仿隐藏了恶意代码的行为表现，但实际上恶意代码本身还暴露在受害主机上，即使是自动化免杀生成的对抗性恶意代码，也可以通过逆向分析来获取恶意载荷，从而能够发现其攻击意图。为了保持攻击的持续性和有效性，攻击意图的隐藏是非常必要的，因此，站在攻击者的角度，安全研究人员预测攻击的一大研究目标应是如何更好地隐藏攻击意图。

当前，我们正处于人工智能新时代的尖端，恶意代码本身及其行为的自动化变种远非人工智能唯一可能的恶意应用。从预测新型赋能攻击技术的角度，2018 年 IBM 研究院的研究人员提出的 DeepLocker[20]便是被预见的一种新型智能恶意代码。DeepLocker 通过主动利用深度学习技术实现了基于黑盒模型的攻击意图隐藏，而对攻击者攻击意图的隐藏便是深度学习赋能效应的体现。

实现攻击意图隐藏的关键点就在于对密钥（用于解密攻击载荷）的隐藏。如图 6 所示，DeepLocker 包含两个主要神经网络模型组件，即目标识别模型和密钥生成模型。如前所述，目标识别模型的核心功能在于精准定位；而密钥生成模型的核心功能则是攻击意图隐藏。具体而言，DeepLocker 将传统以震网蠕虫为例的“if this, then that”的攻击载荷触发条件转换为使用深度学习模型来实现对密钥的隐蔽存储与稳定生成。

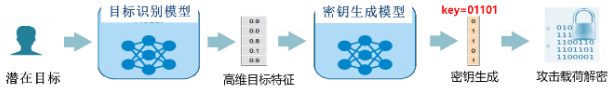

图 6 DeepLocker 攻击意图隐藏示意图

一方面，密钥的隐蔽存储与深度神经网络模型实现的特征识别功能密切相关，目标特征未被检测到时，神经网络模型便不会对正确的密钥做输出。而且由于深度学习模型的“黑盒特性”，使得神经网络模型难以被逆向分析，用于解密攻击载荷的密钥在攻击目标出现之前便可以隐蔽地存储在神经网络模型中。只有当真正的攻击目标出现时，它才会根据高维目标特征生成正确的密钥来解密恶意载荷，从而释放恶意行为。

另一方面，DeepLocker 利用了深度学习模型的泛化性，保证了生成密钥的稳定性。以图 4 所示的人脸识别为例，几乎所有“周星驰”的面部图像在经过目标识别模型和密钥生成模型的处理之后，都将对应完全相同且正确的密钥；而非“周星驰”的面部图像则无法获得正确的密钥。深度学习模型这种对同类事物的泛化性，以及其对不同类事物的异化性，保证了生成密钥的稳定性。

在对 DeepLocker 进行实验验证的基础上，本文深入了解了主动利用深度学习模型做意图隐藏的新型攻击技术，并对其防御难度和可扩展性做了分析与总结。

在防御难度上体现为两个方面：1）DeepLocker 通过使用深度神经网络模型，实现了针对特定目标的精准识别，而防御者由于无法获取特定目标的相关知识，从而难以实现或逆向执行该类恶意攻击的“触发条件”；2）若 DeepLocker 的对抗者想绕过对目标特征的匹配，转而以直接爆破密钥的方式来解锁恶意载荷也是几乎不可能的，因为当 DeepLocker 使用的密钥长度超过了 128 位，破解如此长度的密钥已经被研究者认证为是不可能的事情。即使我们假设安全研究人员得到了 DeepLocker 利用的神经网络模型和加密后的恶意代码，但由于此方面的逆向难度几乎等同于破解 AES 128 位密钥的难度，因此安全分析专家也难以通过逆向分析解密得到有效的攻击载荷，从而无法获知其恶意攻击意图。

在可扩展性上，以 DeepLocker 为例的攻击意图隐藏是以 DNN 作为黑盒模型实现的，从模型扩展迁移的角度，可以基于 RNN、GAN、DRN 等模

型实现更多攻击场景中的意图隐藏。此外，对于只知道输入输出关系而不知道内部结构的系统或设备，如果能够成功实现对攻击意图的隐藏，本文认为都可以将其归类为基于黑盒模型的攻击意图隐藏技术，比如 APT 组织 FIN7 使用的 BIOLOAD 恶意代码，它利用 Hash 黑盒隐藏其攻击意图，便是此类攻击技术的一个具体实例。未来可能会出现很多未知形式的黑盒模型，随着人工智能技术的发展，这些黑盒模型与人工智能的结合，将能够产生更具危害的攻击效果。因此，基于可能或未知的黑盒模型，并研究新型的攻击意图隐藏方式将是未来智能恶意代码发展的一个趋势，需要引起安全研究人员的重视。

3 深度学习助力防御

在深度学习助力防御方面，防御者试图对恶意代码攻击链的各个环节进行逐个击破，通过切断攻击链来达到有效防御的目的。站在被攻击的目标侧，本文探索并总结了三类新型的赋能防御技术，分别为：a.基于深度学习的恶意代码查杀；b.自动化网络钓鱼攻击识别；c.深度学习赋能的恶意行为检测。

值得一提的是，对应图 1，在恶意代码攻击链中，“准备”阶段是攻击者特有的活动，不是防御者可检测或可防御的范畴，因此防御者无法针对该阶段展开赋能防御研究。此外，在“规避”阶段中，如前所述，攻击者会对攻击结果做分析、评估与反馈，为了增强隐蔽生存能力，攻击者会根据反馈信息继续研发具有高隐蔽性的恶意代码，或者将恶意流量伪装变形来规避检测等，以尽可能地降低在攻击活动中被发现的风险，从而在循环反馈中强化攻击效果；而针对以防范规避为目的的新型攻击技术，对应的新型赋能防御技术尚未被开发，本文基于恶意代码攻击链开展综述的主要目的之一便是促进新型赋能防御技术的研究工作。

3.1 基于深度学习的恶意代码查杀

基于深度学习的恶意代码查杀是在“突破”和“存在”阶段提出的防御技术。对应“突破”阶段，防御者一般会在网络边界部署反病毒引擎；到达“存在”阶段，代表攻击者已经在受害主机上建立了立足点，对应的反病毒引擎则应部署在终端主机侧。虽然部署的位置不同，但这两类反病毒引擎对应的查杀对象都是攻击者在“准备”阶段构建的恶意代码。

攻击者构建恶意代码时趋向于使用能够增强其生存性和隐蔽性的技术手段，以达到驻留和持久化的目的。然而，对于一般基于浅层机器学习构建的反病毒引擎，往往难以察觉恶意代码自身潜在、隐蔽性特征的变化。深度学习在提取自主学习并自主提取特征上的优势为防御研究者带来了启发，他们面向攻击链的“突破”和“存在”两个环节，提出使用深度学习技术来学习并提取具有隐蔽性、不可见的特征来表征恶意代码，并对其进行查杀。

通常，基于深度学习的反病毒引擎会将恶意代码作为训练样本，通过对自身的检测系统进行对抗性训练，从而增强恶意代码查杀时的鲁棒性[27]。由于攻击者潜在的对策空间可能很大，单纯的对抗性训练已经被证明可以再次被攻击者利用，因此已有研究者基于对抗性训练，开展了新型的恶意代码查杀技术研究。如图 7 所示，本文将这些新型恶意代码查杀技术总结为三类，分别为增强式对抗性训练、特征压缩、赋能式仿真执行。

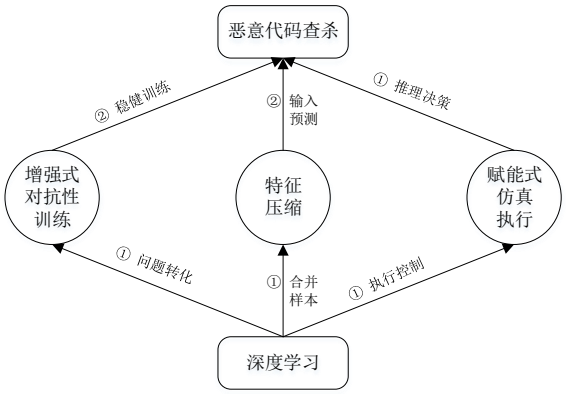

图 7 深度学习赋能的恶意代码查杀技术

1）增强式对抗性训练。2018 年，Abdullah Al-Dujaili[28]等人为了研究减少神经网络恶意代码检测器对抗盲点的方法，他们首先将该问题作为二进制域中的鞍点（即一个维度向上倾斜且另一维度向下倾斜的点）优化问题进行处理。由于鞍点通常被相同误差值的平面所包围，在该平面中，所有维度上梯度值都接近于零，这使得神经网络学习算法陷入其中很难脱离出来，从而降低了神经网络的学习速度。而在利用鞍点优化加快神经网络学习速度的基础上，再对检测系统做稳健训练，可有效减少神经网络的学习盲点。基于此，他们提出了 SLEIPNIR 框架，并且他们在一组 PE 文件①上评估

① https://github.com/ALFA-group/robust-adv-malware-detection

了该方法在面对恶意代码攻击时的鲁棒性，实验证明他们基于对抗性训练的模型检测到的恶意代码样本是自然训练中检测到的恶意代码样本的 4 倍。

2）特征压缩。一种称为特征压缩的新的防御对抗策略在 2018 年被 Weilin Xu[29]等人提出，可以通过检测对抗性示例来强化 DNN 模型。首先，DNN 模型的原始输入空间中，存在许多具有不同关键特征的输入样本，特征压缩将这些样本合并为一个样本，通过“挤出”不必要的输入特征来降低恶意攻击者的自由度，从而减少攻击者可用的搜索空间。然后，该对抗策略的关键思想是将 DNN 模型对原始输入的预测与对压缩输入的预测进行比较，如果原始输入和压缩后的输入产生与模型实质上不同的输出，则输入可能是对抗性的。最后，实验结果也表明通过特征压缩策略可以检测出准确度高且误报少的对抗性示例。

特征压缩策略成功实现的背后，其所依赖的技术原理为：降维表示并不会影响分类器的准确性。具体地，Weilin Xu 等人专注于图像分类，通过减少颜色位深度、空间平滑两种技术手段实现了特征压缩策略。这一策略的实际可行性带来了两个方面的益处：一是有助于提高依赖于图像转化的恶意代码检测引擎[2]的鲁棒性，二是有助于启发对恶意代码行为空间进行特征压缩的研究。

3）赋能式仿真执行。Yu Wang[30]等人于 2019 年提出了一种新颖的基于深度神经网络的恶意控制模型，该模型基于深度强化学习来学习暂停未知文件执行的最佳时间。该模型是第一个基于深度强化学习来保护用户免受恶意攻击的模型，它由执行控制模型和改进的推理模型组成。反恶意代码引擎会模拟执行一个未知文件，并生成一系列行为事件文件。之后，执行控制模型负责处理事件文件并负责控制未知文件的执行；而改进的推理模型则负责决策未知文件属于良性文件或恶意文件。通过实验结果表明，对于测试集中超过 91％的文件，Yu Wang 等人提出的模型比基于启发式的反恶意代码引擎更早地停止执行，能够以更少的延迟提供了更好的保护。

理论上，上述以恶意代码查杀为目的的研究方法中，研究者利用深度学习技术更深入地探索了恶意代码区别于正常代码的独特特征。深度学习在恶意代码查杀上的赋能效应可以总结概括为：通过对基于深度学习的反病毒引擎模型的对抗性训练，可以有效地增强反病毒引擎查杀恶意代码时的鲁棒性。为了进一步了解基于深度学习的恶意代码查杀技术，本文从数据集、检测维度、攻击者策略、模型的可持续性四个方面对该类新型攻击技术进行了剖析，具体如下：

1）数据集。深度学习赋能的反病毒引擎做实际部署的前提是以数据集的数量和质量的保证作为前提的。在实际中，恶意代码的数量是限制反病毒引擎查杀效果的一个重要因素。不仅如此，数据集的质量是另一个影响反病毒查杀引擎的重要因素。攻击者可能会提前发布被污染的数据集，使用被污染的数据集训练后的增强性模型可能携带了不易被发现的后门，攻击者借用后门可以实现其攻击目的。最后，已有的恶意代码数据集可能存在大量的垃圾数据，这也是需要引起安全研究人员关注的问题。所以在针对反病毒引擎做对抗性训练的同时，需要做好数据集的清洗和质量保证。

2）检测维度。如前所述，反病毒引擎的检测维度通常很多，基于深度学习的恶意代码查杀从对恶意代码做检测分类的角度进行了防御诠释，没有考虑对恶意代码载体和恶意代码行为的检测，所以上述理论研究方法中提出的检测引擎目前还不能作为一个完整的反病毒引擎进行实际部署，但作为其中一个检测模块可以用于增强实际中反病毒引擎对恶意代码查杀的性能。尽管如此，本文认为在未来，将恶意代码、恶意代码载体和恶意行为同时作为深度学习模型的输入特征，以实现面向多维度查杀的反病毒引擎是可行的，同时这也需要网络安全研究者投入更多的时间和精力来完成。

3）攻击者策略。未知恶意代码在一定程度是个性化的产物，未知恶意代码表现的攻击者策略与攻击者的心理状态、情绪变化，以及所处的环境等是相互关联的，因此未知恶意代码的产生具有一定的非规律性。基于深度学习的恶意代码查杀是建立在已有恶意代码数据集的基础上的，其对未知恶意代码的查杀受限于攻击者策略的制定。以腾讯 TRP-AI 反病毒引擎为例，作为当前先进的一款基于深度学习的反病毒引擎，TRP-AI 也只是提升了其对恶意代码家族做变种检测的能力，而在面向未知恶意代码的检测上仍存在局限性。

4）模型的可持续性。恶意代码的数据分布规律往往不具有长期稳定性，这使得恶意代码攻防成为了一个不断对抗与制衡的过程，因此深度学习赋能的反病毒引擎往往也不具有可持续性，即面临着模型的更新与退化问题。Lorenzo Cavalaro 团队在

USENIX Security2017[31]和 USENIX Security2019[32]发表的论文中指出，模型的更新、退化问题与恶意代码数据在时间或空间维度上的概念漂移（Concept Drift）密切相关。

具体地，在时间维度上，攻击者可能会利用“恶意代码的查杀模型发布”与“新型恶意代码出现”之间的时间差，用“未知”的恶意代码构造攻击，以逃过深度学习模型的检测；在空间维度上，攻击者可能会利用“训练集”与“测试集”之间数据分布的不同，用不同分布的恶意代码构造攻击来实现免杀。时空维度相辅相成，时间维度上的概念漂移必然是由于空间维度的分布不同引起的，而空间维度的概念漂移也必然会导致时间差的出现，两个维度相互作用，将“概念漂移”实例化为多种不同的赋能攻击技术，不断挑战反病毒引擎的防御边界。

而对于安全防御人员来说，为了防止模型的退化，或者为了及时对模型进行更新，在利用深度学习开展恶意代码查杀等防御研究时，不论从空间或时间上，都不能停止对模型的更新。模型在训练与测试中的“时间一致性”和“分布一致性”应是保证恶意代码查杀鲁棒性的重要准则。

综上所述，对于基于深度学习的恶意代码查杀，很大程度上需要安全厂商的未雨绸缪，通过利用已获取的恶意代码或变种样本，安全厂商可在保证“时间一致性”和“分布一致性”的基础上，提前强化和攻击自己的反病毒引擎模型，防止因恶意代码数据的“概念漂移”导致模型退化，在发现自身模型的“盲点”时，通过及时修复盲点，达到提升防御能力的目的。而在自我攻击与模型强化的过程中，安全厂商则需要站在攻击者的角度，合理预测攻击者策略，并在保证数据集的数量与质量的基础上，才能更好的发现模型“盲点”，达到多维度检测的效果。

3.2 自动化网络钓鱼攻击识别

自动化网络钓鱼攻击识别是深度学习在“投递”阶段进行赋能防御的产物，该阶段的检测对象是恶意代码投放过程中的传播媒介，如钓鱼邮件和社交帖子等。由于当前传统的网络钓鱼已经被证明可以利用黑名单、启发式、机器学习等方法实现成功对抗，攻击者又展开了对新型自动化网络钓鱼攻击技术（比如基于 NLG 的自动化网络钓鱼）的探索。由于 Twitter、Facebook 等社交网站已经成为攻击者展开攻击的重灾区，而人工智能技术的发展也使得大规模自动化网络钓鱼活动成为可能，由此，本文重点关注对自动化网络钓鱼攻击的识别技术，并以此为切入点展开防御。如前所述，NLG 技术被用于新型自动化网络钓鱼攻击，同样地，从安全防御的角度，基于深度学习的 NLG 技术也可以被用于从语义层实现对自动化网络钓鱼的检测。

对于深度学习赋能的钓鱼攻击识别，当前的理论研究相对较少，比较典型的一项研究是 Sneha Kudugunta[33]等人在 2018 年提出的面向社交推文级别的检测方法。Sneha Kudugunta 等人强调数据的可解释性，通过基于文本特征和用户元数据的组合设计 LSTM 深度神经网络构建了第一个基于推文的社交机器人检测系统。由于 LSTM 能够学习复杂的非线性特征，在训练好的 LSTM 模型中，大多数隐藏单元在真实推文和 bot 推文的激活值分布上有着显著的差异。通过实验证明，仅从一条推文中，该检测系统就可以在将社交机器人与人类分离的过程中实现了 96%的分类精度；当将相同架构应用到账户级别的检测中时，他们的方法也实现了高达 99%的分类准确性。另外在 2020 年，Gan Kim Soon[34]等人利用深度神经网络对网络钓鱼电子邮件进行了检测，并且通过实验证明 DNN 在钓鱼邮件的检测上可以实现 94.27%的检测性能，进一步证明了深度学习技术在自动化网络钓鱼识别中的可行性。

在自动化网络钓鱼识别领域中，深度学习的具体赋能效应体现在对网络钓鱼推文或钓鱼邮件的检测性能上。然而，机器学习也可以在该领域中实现高的检测效果，因此，进一步地，本文对比分析了机器学习与深度学习在钓鱼检测中的性能，并探索、总结了反钓鱼网络研究的难点问题及其未来的发展趋势。

1）机器学习 VS 深度学习：

对于以钓鱼邮件和社交媒体帖子为代表的恶意代码传播载体来说，相比以二进制文件、PE 文件等多种形式呈现的恶意代码，这些待检测载体多以短文本的形式存在，基于机器学习的方法非常适合对这类数据进行恶意检测。由于人工选择的特征已经足以满足对短文本内容的检测，所以在深度学习出现之前，基于机器学习的自动化钓鱼检测技术已经趋于成熟。因此，尽管上述研究结果证明了基于深度学习的钓鱼攻击识别的显著性能，但对于这类短文本数据，基于机器学习的检测仍然属于主流检测技术。2020 年，Gan Kim Soon[34]等人对比了机器学习中的浅层神经网络与深度学习中的深度神经网络在网络钓鱼检测中的性能，如表 4 所示的对比

结果显示，使用机器学习的检测效果仍要略高于基于深度学习的检测效果。

表 4 机器学习与深度学习在钓鱼检测上的性能对比

神经网络	平均检测准确率
浅层神经网络（机器学习）	94.41%
深度神经网络（深度学习）	94.27%

然而，未来网络钓鱼载体中是否存在被攻击者利用且具有较高隐蔽性的特征仍尚未可知，所以在以机器学习为主导的恶意代码传播载体的检测中，本文认为深度学习赋能的自动化网络钓鱼识别仍具有研究意义和价值。未来机器学习与深度学习的融合或许能够为自动化网络钓鱼识别带来更好的检测性能。

2）难点问题：

自动化网络钓鱼攻击识别尤其强调数据集的获取，依赖大量样本做训练的深度学习检测模型才能更好地发挥它们的优势与作用。上述钓鱼检测方法使用的数据集，是由 Sneha Kudugunta[33]等人通过 SMOTE 过采样从最少的标记数据中生成适合深度神经网络训练的大型标记数据。另外，为了深入了解攻击原理及增强检测性能，安全研究者也需要实际的攻击样本，由 Norah Abokhodair[35]团队和 Juan Echeverria[36]团队分别于 2015 年和 2017 年偶然发现的两个大型社交僵尸网络便是稀有且珍贵的真实数据集。

但由于研究人员不太可能一直有如此运气获得大量真实的攻击样本，而且为了保护用户隐私，有效的钓鱼检测将更多地依赖于网站服务提供者，如 Twitter、Facebook 等的运营商。因此，如何更好地弥合研究者与运营商之间的研究差距，以及如何在保护数据私密的前提下，建立两者之间的共享数据集将是当前亟待解决的一个难点问题，也是促进自动化网络钓鱼识别研究的一大需求。

3）发展趋势：

Sneha Kudugunta[33]等人提出的推文级检测方法是基于 Twitter 网站提出的。本文关注 Twitter 网站使用的 HTTPS 协议，并通过深入探索得出结论：未来网络钓鱼攻击的一大发展趋势就是把攻击建立在 HTTPS 协议上。

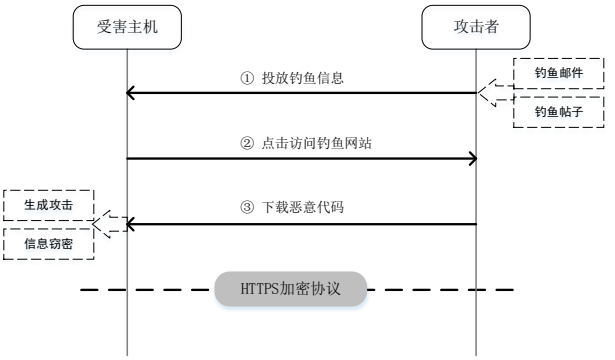

图 8 基于 HTTPS 的钓鱼攻击流程

这一结论的得出基于两大原因，一是基于网络流量的分析，如图 8 所示的钓鱼攻击流程，它从网络流量传输的角度展示了基于 HTTPS 的恶意代码投放过程。投放钓鱼信息、点击访问钓鱼网站、下载恶意代码三个步骤是基于 HTTPS 加密协议完成的，而 HTTPS 的使用将有助于恶意代码避开浏览器的安全检测，并可能会给网络访问者带来错误的安全感，无论从技术或心理因素上都将能够帮助提高钓鱼攻击的成功率；二是数据统计的结果，根据 APWG（Anti-Phishing Working Group）的报告[37]，HTTPS 的使用呈逐年上升趋势，尤其在 2016 到 2017 一年的时间内，所有网络钓鱼站点中 HTTPS 的使用比率从 5%上升到了超过 31%。

综上，当前针对自动化网络钓鱼攻击的识别多通过获取社交网站运营商的 API 或在取得目标用户邮箱权限的情况下进行数据集的爬取，并基于文本内容来检测钓鱼攻击的存在。在内容检测之外，未来网络钓鱼攻击的发展趋势也启发安全研究人员关注到另一个比较重要的检测视角，即基于流量的检测，尤其是基于 HTTPS 流量的恶意代码传输。基于流量的检测是一种重要的防御手段，本文将其纳入深度学习赋能的恶意行为检测这类技术中，将在下节中对其进行介绍。

3.3 深度学习赋能的恶意行为检测

对于防御者而言，恶意行为检测是比较直接的一种防御手段，防御者可以在“影响”和“C2”阶段中部署恶意行为检测。（1）“影响”阶段中，在受害主机上建立了立足点的恶意代码会释放恶意活动（比如破坏数据、破坏硬件、窃取密码等），防御者可以通过监控主机进程行为等方式，利用深度学习对恶意进程行为进行赋能检测。（2）在“C2”阶段，出于对持久化和循环控制的需要，攻击者会利用命令与控制信道做命令的更新传达，受害主机也会在该信道上对窃密信息进行回传，安全防御人

员可以对命令与控制信道上的往返流量继续检测和过滤。攻击流量是恶意行为的一种表现，因此本文将“恶意流量检测”和针对终端主机的“进程行为检测”统称为“恶意行为检测”。

- **恶意流量检测**

对网络流量的分析是对在受害主机上运行的分散式反病毒引擎的补充。它使网络管理者可以在整个网络中一致地实施安全策略，并尽可能地减少管理开销。当前的很多反病毒引擎借助机器学习的优势来构建恶意流量检测系统，并结合专家知识，通过手工定制检测特征以期实现高精度的检测性能。但随着恶意代码的不断发展变换，其产生的网络流量数据中所携带的特征也是不断变化的，因此基于专家知识生成的流量特征可能会失效，其针对不同的问题和情况也无法迁移使用。为了减少甚至摆脱对专家知识的依赖，研究人员提出利用深度学习来赋能流量检测，而深度学习在恶意流量检测方面的赋能效应体现为通过自动特征提取而实现的对恶意流量的检测性能上。

2018 年 Sajad Homayoun[38]等人提出了一种利用深度学习且独立于底层僵尸网络体系结构的恶意网络流量检测器，称为 BoTShark（Botnet Traffic Shark）。BoTShark 采用两种深度学习检测模型（堆叠式自动编码器 Autoencoder 和卷积神经网络 CNN），以消除检测系统对网络流量主要特征的依赖性。BoTShark 的优势在于它借助神经网络可以自动提取特征而无需专家知识。Sajad Homayoun 等人使用具有通用性、现实性和代表性的僵尸网络数据集 ISCX①，并通过实验表明，BoTShark 能够从两种常见的僵尸网络拓扑（即集中式和 P2P）中检测僵尸网络流量。在检测僵尸网络的恶意流量时，BoTShark 实现了 91%的分类准确率和 13%的召回率。

2019 年 Gonzalo Mar ń[39]等使用了与知识完全无关的输入（仅原始字节流）来探索深度学习模型在检测和分类恶意流量的强大功能。Gonzalo Mar ń 等人使用包含一百万个样本的数据集，并考虑两种原始的数据表示形式：数据包和流。通过实验评估深度学习技术在恶意软件和网络攻击检测任务上的性能，特别在使用原始数据流作为输入情况下，Gonzalo Mar ń 等人使用的深度神经网络模型展示了出色的性能：（1）提供了高度准确且实际可行的恶意软件检测结果；（2）比基于随机森林的浅层模型能更好地捕获基础恶意软件和正常流量；（3）与通过领域专家基于知识的检测器获得的检测效果一样好，而无需任何手工特征。

除此之外，在流量检测中也需要特别注意对 HTTPS 流量的检测。如图 9 所示的基于 HTTPS 协议的加密通信流程，通过使用加密的 HTTPS 协议可以很容易地防止对 HTTP 有效负载的分析，网络攻击者很容易利用 HTTPS 隐藏攻击意图，而且 Google，Facebook，LinkedIn 和许多其他流行的站点默认情况下都使用基于 HTTPS 的加密网络流量，这使得基于网络或社交媒体的恶意攻击更具优势，所以随着 HTTPS 使用量的增长，恶意流量分析也有必要考虑对 HTTPS 的检测。

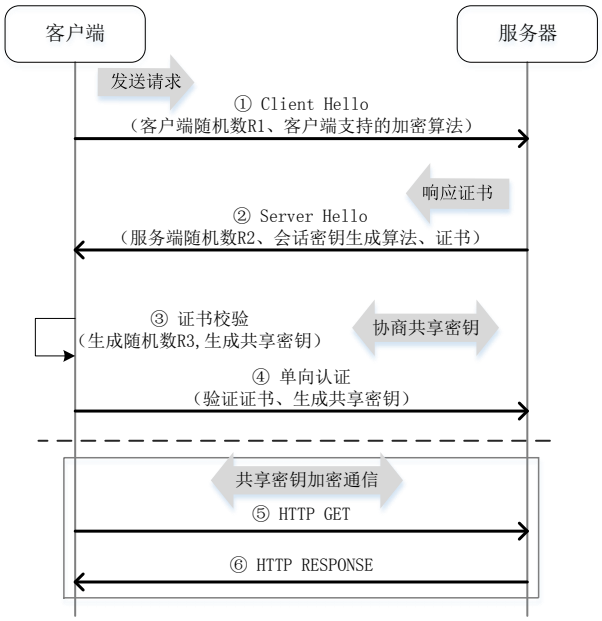

图 9 HTTPS 通信流程

一个典型的 HTTPS 流量检测的研究成果是通过分析基于 HTTPS 的加密通信流程(如图 9 所示)，从而在 2017 年被 Paul Prasse[40]等人提出的。Paul Prasse 等人在无法通过对有效传输载荷提取特征的情况下，开发并研究了基于 LSTM 的加密恶意流量检测模型，该模型在经过为期两个月的训练之后，可以对许多不同的恶意软件家族识别其未知的加密恶意流量。

综上，在面向流量的赋能检测中，HTTPS 协议相比 HTTP 协议能更好地实现对恶意流量的隐蔽，因此攻击者越来越多地使用加密流量来掩盖网络威胁。据 Gartner 报告，2019 年进行的网络攻击中有 60%利用加密，而到 2020 年，这一数字将上升到 70%[41]。因此，本文尤其关注基于 HTTPS 的加

① https://www.unb.ca/cic/datasets/index.html#Botnet%20Data%20set

密恶意流量检测。以 Paul Prasse[40]等人提出的检测模型为实例，本文对其工作原理进行了深入探究，并对分析了当前面向加密恶意流量检测的实际可行性。

面向加密恶意流量检测的工作机理与 HTTPS 的通信流程密切相关。具体地，一般网络设备会将在一对 IP 地址和端口之间交换的 TCP / IP 数据包做聚合，其中相关的地址、时间和数据量信息会被保存到日志文件中。大多数情况下，网络流量分析者可以看到未加密的主机域名，即受害主机向 C&C 服务器发送 HTTPS 请求时（对应图 9 中的①步骤），包含在 Client Hello 数据包中的主机域名是可见的。除此之外，往返于受害主机与 C&C 服务器之间的恶意流量的时序以及数据量是可统计的。Paul Prasse 等人通过对可见域名和统计信息的赋能分析，实现了有效的加密流量检测。Paul Prasse 等人的方法通过对流量侧攻击数据的分析，在一定程度上补充并完备了深度学习赋能的恶意流量检测手段。但加密恶意流量检测的研究仍相对较少，且当前各大安全厂商也没有推出利用深度学习来赋能加密恶意流量检测的安全产品，因此，基于加密恶意流量的理论研究仍尚待完善；而面对攻击者基于加密流量展开的攻击趋势，在该领域中也有待提出更有效且实际可行的加密流量赋能检测方案。

- **进程行为检测**

面向“C2”信道的恶意流量检测是面向网络侧进行防御部署的有效手段，很多部署在网络边界的流量检测系统（比如 IDS、IPS 等）专注于恶意流量的检测和过滤。但很多高级的未知恶意软件还会利用深度学习赋能来模仿良性流量使得攻击变得隐蔽。因此，在恶意行为检测上，单纯基于流量的恶意软件赋能检测变得十分困难，还需要于主机端的行为检测相结合。

事实上，安全防御人员不仅可以分析流量，还可以通过行为分析来监控主机上是否被感染了恶意代码。对此，研究者也提出很多可行的技术手段。例如，2016 年，Shun Tobiyama[42]等人提出了一种新的恶意软件进程检测方法，在假设可以使用语言模型来提取进程行为特征的基础上，该方法首先监视进程，将 API 调用序列记录为进程行为并生成日志文件，然后将 RNN 与 CNN 相结合，使用 RNN 来提取进程行为的特征，并使用 CNN 对特征图像进行分类。Shun Tobiyama 等人以 81 个恶意软件进程日志文件和 69 个良性进程日志文作为数据集，在 Cuckoo 沙箱评估了他们的方法，当特征图像尺寸为 30×30 时，该方法实现的最佳检测准确性达到了 96%。

2018 年，Matilda Rhode[43]等人首次提出了一种早期行为检测模型，该模型能够在执行文件的前几秒钟中预测文件是否为恶意文件。该检测模型使用 RNN 进行开发，并且由于恶意软件的快速发展，新发现的样本应及时用于 RNN 的训练，对应 RNN 的模型架构也有修改的需要，因此 Matilda Rhode 等人通过对超参数空间进行随机搜索，通过发现模型的良好配置，进而实现了模型训练与构建的自动化。以 3000 个勒索软件样本作为数据集，实验结果表明，Matilda Rhode 等人使用一组 RNN，仅用 5s 的动态数据就实现了 94％的检测精度，而在不到 10s 的时间内实现了 96％的检测精度。

通过对上述面向进程行为做检测的研究论述，深度学习的在该领域的赋能效应表现为对进程行为数据的检测精度上。对此，本文分析认为，面向深度学习的进程行为检测在理论上是可行的技术手段。具体原因有两点：

1）恶意代码在攻击过程中，无法避免地会留下行为足迹，这些足迹的发现就需要面向终端主机的行为检测。而对于实现了隐蔽性的恶意行为，传统基于规则、基于机器学习的方法很难捕获相关特征并发现恶意行为。深度学习因具有自动提取深层特征的能力，因而能够对恶意进程行为实现较高的检测精度。

2）主机进程行为生成的日志文件或指标快照是可被获取的大量数据，一些恶意的、隐蔽的行为特征被包含其中。而大量可获取的样本则是深度学习类检测模型做健壮性训练的良好基础，有利于发掘具有隐蔽性的恶意特征，促进快速检测和恢复。

表 5 两种进程行为赋能检测方法对比

文献	进程行为捕获渠道	行为指标/特征	记录时间	检测率	问题/不足
[42]	日志文件 （记录 API 调用）	①Process Name ②PID ③Event ④Path ⑤Result ⑥Detail	5min * 10 次	96%	1）API 调用容易受到操纵，会导致神经网络对样本进行错误分类；

					2）进程行为捕获时间过长。
[43]	指标快照（记录机器活动特征）	①system CPU usage ②user CPU use ③packets sent ④packets received ⑤bytes sent ⑥bytes received ⑦memory use ⑧swap use ⑨total processes ⑩maximum process ID	20s（每秒 1 次快照）	96%	对在开头植入长时间睡眠或良性行为的恶意代码，无法实现检测。

基于理论层面的分析，本文对深度学习赋能进程行为检测的实际可行性也做了深入分析与总结。本文分析对比了上述两种方法的技术细节，通过提炼这些方法当前存在的问题或不足，启发对实际可行性的思考，旨在为之后开发更完善的进程行为赋能检测技术提供依据。

如表 5 所示，从行为特征分析，Shun Tobiyama[42]等人和 Matilda Rhode[43]等人分别通过记录 API 调用序列和记录机器活动特征来实现对进程行为的检测，由于 API 调用容易受到恶意攻击者的操纵，这会导致利用神经网络的检测出现分类错误，攻击者将能够绕过面向 API 的行为检测；相比之下，有关机器活动的行为特征无法被攻击者篡改，而神经网络能够通过自主学习，提取出区别于恶意进程行为与良性进程行为的深层次、不可见的特征，因此基于机器活动的恶意进程行为检测在检测性能的鲁棒性方面是优于基于 API 调用的检测方法的。

从记录时间分析，Shun Tobiyama[42]等人和 Matilda Rhode[43]等人分别在文件执行期间以分钟级和秒级开展对进程行为的检测。从对恶意行为做及时捕获的角度，Shun Tobiyama[42]等人以 5 分钟为周期做行为捕获，其捕获时间相对较长，意味着恶意代码在被检测到之前可能已经释放了攻击动作，甚至造成破坏，从而导致错过阻止恶意行为的机会；相比之下，Matilda Rhode[43]等人提出的秒级早期行为检测方法使网络安全终结点保护功能得以增强，即可以及时有效地阻止恶意代码的攻击行为，而不是在执行后检测并做修复，从理论上分析，这种早期行为检测的实现似乎为恶意行为检测提供了比较可观的时效性。但从另一个角度分析，像 APT 这样的实际攻击，其特性就是跨地域、时间长、不连续，而这对于安全防御来说，即使耗时数周甚至数月才能检测出其对应的恶意行为，也是非常实用和有价值的。因此，本文结合对实际攻击和检测案例的分析，得出结论：在实际的攻防对抗中，追求分钟级甚至秒级的防御检测往往是不可行的。

综上，通过两个维度的分析对比，本文认为当前面向主机进程行为的实际部署尚有难度。但上述两种基于进程行为的赋能检测方法各有利弊，在不考虑检测耗时的情况下，可尝试优势叠加效应，即结合两者的优势强化检测性能，并进一步增加行为检测在实际应用中的可行性。目前本文尚未对其进行深入的理论剖析和实验验证，在未来的研究工作中将以此为启发点继续深入研究，并探索更多面向实际应用的赋能行为检测技术。

3.4 基于攻击链的赋能防御技术分析与总结

由图 1 所示的恶意代码的一般攻击过程，以及本文对赋能攻击技术的讨论，可以得出结论：恶意代码攻击是分阶段、按计划、隐蔽地、逐步推进的一个过程。从攻防博弈的角度，如能根据滑动标尺模型，围绕恶意代码攻击过程建立纵深防御体系，覆盖攻击的主要环节，则能有效地检测和拦截攻击。因此，对应恶意代码攻击链，本文提取了三类主要的赋能防御技术，这三类防御技术分别面向恶意代码攻击的“投递”、“突破”、“存在”、“影响”和“C2”五个环节展开检测，基本覆盖了攻击的主要环节。在这些主要攻击环节上应用深度学习来实现赋能防御，体现在深度学习的赋能效应上则是大大提升了安全防御侧的网络威胁发现能力。

表 6 深度学习助力防御技术对比与分析

分类	特点/技术	年份	文献	前提/假设	验证引擎	方法描述
基于深度学习的恶意代	增强式对抗性训练	2018	[28]	白盒假设	DNN	通过鞍点公式将模型学习视为一个鲁棒的优化问题

码查杀	特征压缩	2018	[29]	1）白盒假设； 2）攻击者未意识到特征压缩的使用。	DNN	特征压缩(将与原始空间中许多不同特征向量相对应的样本合并为一个样本，从而减少攻击者可用的搜索空间)
	赋能式仿真执行	2019	[30]	1）白盒假设； 2）计算机之前尚未受到感染； 3）在无网络访问的反恶意软件引擎的仿真器中观察到的行为与在用户计算机的本机操作系统上执行该文件时观察到的行为类似。	DRN	学习暂停文件执行的最佳时间
自动化网络钓鱼攻击识别	面向社交网站	2018	[33]	获取对 Twitter API 的使用权限。	LSTM	提出基于 LSTM 的深度神经网络架构，利用内容和元数据在推文级别检测 bot
	面向钓鱼邮件	2020	[34]	白盒假设	DNN	利用 DNN 网络进行钓鱼邮件检测
深度学习赋能的恶意行为检测	恶意流量检测	2018	[38]	假定隐藏层必须以尽可能少的失真量重建初始信息（白盒假设）	CNN	基于深度学习自动提取流量特征，独立于僵尸网络拓扑检测来恶意流量
		2019	[39]	白盒假设	CNN LSTM	提出并评估了不同的 DL 架构和不同的、与知识无关的输入表示形式在恶意流量检测上的性能
	恶意进程行为检测	2016	[42]	1）白盒假设； 2）假设可以使用语言模型来提取进程行为的特征，并可以用进程行为的特征去训练 RNN。	CNN+RNN(LSTM)	通过记录日志文件，使用深度神经网络对恶意进程进行分类
		2018	[43]	1）白盒假设； 2）恶意行为会留下可测量的足迹； 3）一旦恶意文件开始执行，恶意活动就会迅速开始。	RNN	根据行为数据的简短快照来预测可执行文件是否为恶意软件

- **技术层面的分析**

本文在对这三类技术进行了详细介绍之后，进一步对它们进行了归纳与总结，旨在为智能化网络安全防御人员提供系统化的技术链框架，促进防御技术的发展。根据表 6 所示，本文做出如下四点思考与总结：

第一，深度学习赋能的防御技术与攻击技术最为明显的区别在于对研究前提或假设的不同设定。对于攻击技术的预测，研究人员力图做到在最苛刻的条件（即黑盒假设）下验证攻击技术的性能；而在对防御技术的开发中，研究人员则通过白盒假设不断改造和完善检测模型，以追求最佳的检测效果。对攻击者来说，攻击者在实施攻击时往往很难知道反病毒引擎的模型结构和内部参数等信息，为了绕过反病毒引擎的查杀，建议的方法就是在实际攻击之前，基于黑盒假设预先验证攻击技术的可行性，降低被查杀的风险。而对于防御者而言，强化后的检测模型如果实际可行，将被作为反病毒引擎的一个模块进行实际部署，并对恶意代码进行查杀，因此在实际投入使用之前研究者往往会通过白盒假设进行模型提炼。

第二，在大多数情况下，攻击者和防御者面向的目标对象都是反病毒查杀引擎，因此在赋能防御的研究中，研究人员使用的验证引擎也是涵盖了 DNN、RNN、DRN 等多种模型架构，具体如表 6 所示。但对于这些模型的部署，本文通过攻防对抗的分析，给出如下建议：人工智能的相关技术同样可以为攻击者所使用，为了降低被攻击者反向利用的机会，反病毒引擎的实际部署和检测中切记不要暴露对恶意文件的真实评分情况。这是因为评分一旦给出，相当于间接向攻击者呈现了攻击效果，这对攻击者指定下一步攻击策略具有一定的指导作用。

第三，深度学习技术赋能的防御技术不是万能的，具体体现为两点：1）当前很多防御技术的研究方法，其实际可行性并不高；2）以自动化网络钓鱼识别为例，当前机器学习的检测性能要高于深度学习的检测效果。因此，未来机器学习与深度学

习结合的可行性有待被研究，而深度学习技术如何在安全防御发挥其更大的效能也有待被进一步发掘。

第四，除了对赋能检测模型的设计与强化，一系列的安全防御措施也应予以加强。一方面，网站建设者和各大服务运行商也应加强管理与规范，比如管控相关 API 接口的使用等，降低钓鱼攻击的发生概率；另一方面，终端用户也应提高个人安全意识，注意保护并防止个人隐私信息泄露，并时刻保持警惕性。

基于以上思考，并结合上述对深度学习助力防御技术的分析，可以看出以深度学习为代表的人工智能技术带来了防御上的一定优势，但这也会给攻击者提供一个新的攻击面。如上所述，基于深度学习构建的防御工具不是万能药，研究人员预测攻击者可能正在对基于深度学习的防御保护中寻找弱点，并通过使用人工智能技术设计低成本、更有效的攻击。然而这类攻击区别于人工智能的赋能效应，属于人工智能的伴生安全问题，本文将在第四章节对其展开介绍。

- **管理层面的分析**

深度学习作为一种能助力攻击的手段和工具，大幅提升了攻击能力，但单纯依赖技术往往并不能有效解决问题。因此，还必须通过法律法规、政策制度、标准规范等管理举措来提升防御水平。

以 DeepFake[①]事件为例，DeepFake 借助深度学习将图像或视频中的人物肖像进行替换，其在名人色情视频、复仇色情、虚假新闻、恶作剧和金融欺诈中的应用引起了全世界的广泛关注，中国、美国、英国、加拿大等国家纷纷指定了一系列法律、政策、标准等安全管理举措以发现并限制对 DeepFake 技术的滥用。

从 DeepFake 事件也可以看出深度学习技术一旦被攻击者利用，将带来巨大的安全隐患。例如：一个绝对安全的深度学习产品，也无法防止被别有用心者恶意利用，法律法规则可以有效威慑这种恶意利用行为，保护深度学习助力安全的发展；一个未必安全的深度学习产品，如果没有经过严格的安全评测就进入市场必然会带来安全问题，政策制度则可以强制设置市场准入条件；安全合规的深度学习产品，如果没有按照一致的标准规范进行设计，就如同计算机都不按照统一的 TCP/IP 协议联网，必然导致互联网无法运转一样，有标准规范的制约才能保证深度学习助力安全的生态环境良性发展。

因此，随着深度学习在恶意代码攻防对抗中的不断应用，对深度学习助力安全领域的法律、政策、标准等的制定与完善是十分必要的，这些管理举措的实施尽管不会杜绝攻击，但却可使其攻击成本显著提升。

具体而言，在法律法规层面，需要在相关立法中确立人类优先的原则、制定相关的法律条例、限制深度学习技术被恶意利用等；在政策制度层面，需出台政策明确深度学习产品设计指南、伦理审查办法、安全监管办法、重要场景下智能算法公平性/透明性/可解释性规定、安全问题可追责制度、数据来源保护条例、隐私数据保护条例等；在标准规范层面，深度学习产品及零配件的制造商需遵守深度学习助力安全的相关标准和规范来研发智能系统，减少深度学习产品漏洞，遵守产品安全评测标准，统一深度学习系统间的通信协议，从而提升系统的安全性和兼容性。当然，以上提到的深度学习相关管理举措，应该也适用于机器学习、强化学习、联邦学习等更广泛的人工智能技术。

4 深度学习的伴生安全问题

深度学习作为人工智能领域的一项新技术，既能赋能安全，又会有伴生安全问题（又称之为“安全伴生效应”）的产生。所谓安全伴生效应，是指尽管新技术在酝酿之初会根据以往的经验去充分地考虑安全问题，但在新技术推出之后，势必会在不断的应用中发现新的安全问题伴生而来。如前所述，伴生安全效应表现在两个方面：一是新技术的脆弱性导致新技术系统自身出现问题，无法达到预设的功能目标，我们称之为新技术自身带来的“内生安全”；二是新技术的脆弱性并没有给新技术系统自身的运行带来什么风险，但这些脆弱性却可以被攻击者所利用，从而引发其他领域的安全问题，我们称之为“衍生安全”。

当前，深度学习被越来越多地应用于网络攻防安全对抗中，而由深度学习系统自身脆弱性导致的衍生安全问题也越来越多地受到了安全研究者的关注[44, 45]。其中，模型后门攻击与防御是与当前恶意代码的赋能攻防对抗紧密相关的研究，在“突破”阶段也对恶意代码的攻击与防御起到了助力作用，因此，本文将其纳入综述范围，在本节中将对它们

① https://en.wikipedia.org/wiki/Deepfake

展开介绍。

4.1 深度学习伴生的模型后门攻击

后门（错误的数据或恶意行为等）是模型设计者不关注的一种输入类型，在保证模型正常运行的前提下，攻击者将后门注入神经网络模型的训练中，让模型学习到非预期的内容，从而影响模型决策的方式称为模型后门攻击，或者特洛伊木马后门攻击。这种被植入后门的模型，在面对正常输入时依旧可正常输出，但面对带有后门触发器（又称为木马触发器，特洛伊木马触发器）的输入时，会将其识别为攻击者想要伪装成的对象，即将其错误分类为攻击者指定的目标标签。

在可预见的未来，公开发布的深度学习模型将会变得流行，训练好的深度学习模型将会像日常商品一样成为消费品，这些模型可以用于发布、共享、再训练或转售，这就为攻击者提供了很多进行模型后门攻击的机会。攻击者可以下载这些公共的深度学习模型，并在对其植入后门之后，重新发布和共享模型。由于深度学习具有迁移性，一旦这些携带后门的深度学习模型被迁移应用在各种场景中，则会带来巨大的安全隐患。

为了深入了解深度学习模型自身存在的脆弱性，研究学者已经开展了许多针对模型后门攻击的预测性研究，以期站在攻击方的角度研究前沿技术，掌握在模型后门攻防中的主动性。例如，2017 年，Yingqi Liu[46]等人在假设攻击者对目标神经网络具有完全访问权限，但却无法访问原有的训练数据和测试数据的基础上，使用外部数据重新训练模型，仅在几分钟到几个小时的时间内便可在后门触发器和所选神经元之间建立强连接，且不影响神经网络模型的正常功能，实现了高达 97.15%的攻击成功率。同年，Yujie Ji[47]等人首次证明了存在潜在危害的原始学习模块（Primitive Learning Module，简称 PLM）会对由其“组合”生成的深度学习系统造成严重的安全威胁，即被植入了后门的 PLM 具有语法和语义的不可辨性，几乎与其良性版本无法区别，但当存在携带后门触发器的输入时，恶意 PLM 便能够强制深度学习系统发生故障，并在对 DNN 模型的攻击验证中获得了 95%的成功率。也是在 2017 年，Dawn Song[48]团队在假设攻击者不能直接访问目标学习系统，且只能向训练数据中注入少量后门样本的前提下，研究并提出了两种物理上可实现（比如利用佩戴眼镜的不同）的后门攻击——输入实例密钥攻击和模式密钥攻击，并且通过评估证明攻击者只需注入大约 50 个后门样本，就可以实现超过 90%的攻击成功率，成为了第一个证明可以创建物理上可实现的后门攻击而不涉及模型训练过程的工作。

2018 年，Cong Liao[49]等人专注于深度学习中的 CNN 模型，不仅基于各种假设对模型后门攻击进行了广泛评估，而且分别在模型训练之前和模型更新期间执行后门注入，这是唯一一项考虑不同程度的攻击者知识和能力的工作，并且经过攻击验证，即使在最苛刻的假设下，攻击者也能够实现 90%的攻击成功率。2019 年 Adnan Siraj Rakin[50]等人基于对 DNN 模型的白盒假设首次提出了一种基于比特翻转的模型后门攻击方案，即 TBT（Targeted Bit Trojan）。攻击者通过翻转神经网络模型中易受攻击的权重位，将功能完备的 DNN 模型转换为木马后门模型，消除了模型重新训练的需要，并在基于 CIFAR10 数据集的攻击验证中，实现了 93%的攻击成功率。

上述模型后门攻击方法是分别基于不同的假设、数据集对各种不同的深度学习模型展开的攻击，并且这些方法分别从不同的维度对模型后门攻击进行了预测，旨在促进神深度学习模型的鲁棒性研究工作，为防御机制的发展做出贡献。如表 7 所示，为了更直观地呈现不同方法间的特点，本文将它们进行了归纳与总结。

表 7 深度学习伴生的模型后门攻击技术对比与分析

年份	文献	前提/假设	攻击引擎	数据集（来源）	攻击成功率	技术贡献
2017	[46]	白盒假设（攻击者对目标神经网络具有完全访问权限），但却无法访问原有的训练和测试数据	DNN、CNN (共 5 个)	原始数据集[51-54] 外部数据集[55-57]	96.58% 97.15%	仅在几分钟到几个小时的时间内便可在触发器和所选神经元之间建立强连接。

2017	[47]	黑盒假设，但攻击者拥有对训练数据集的访问权限	DNN (数字皮肤癌筛查系统)	the International Skin Imaging Collaboration (ISIC) Archive	95%	首次证明了潜在有害的 PLM 会对组合生成的深度学习系统的安全性产生巨大威胁，使它们远离预期的行为。
2017	[48]	弱威胁模型假设： 1）黑盒假设，且攻击者不了解其训练数据集； 2）只允许攻击者向训练数据中注入少量的后门样本	DNN (DeepID、 VGG-Face)	YouTube Aligned Face datase	>90%	第一个显示弱威胁模型下后门攻击可行性的工作; 第一个证明模型后门攻击可以创建物理上可实现的后门而不涉及训练阶段的工作。
2018	[49]	三种不同的假设： 1.白盒假设，且攻击者具有完整的训练数据知识； 2.攻击者要么只具有模型体系结构的知识（黑盒假设），要么只能访问训练数据（白盒假设）； 3.既是黑盒假设，攻击者也不了解训练数据。	CNN	GTSRB (German Traffic Sign) 、MNIST、CIFAR-10	>90%	唯一一项考虑不同程度的攻击者知识和能力的工作。
2019	[50]	白盒假设	DNN (VGG-16、 Resnet-18)	CIFAR10、SVHN、ImageNet datasets	93%	翻转易受攻击的 DNN 权重位，不需要重新进行模型训练。

首先，与基于对抗样本生成的自动化免杀技术的目的相同，模型后门攻击也被用于在恶意代码攻击的生命周期内做生存对抗，而且都是通过深度学习模型的误分类来完成的，所以有时不可避免地，研究者会将两者混淆，因此，本文将基于对抗样本生成的自动化免杀与深度学习伴生的模型后门攻击进行了区别化分析，具体如表 8 所示。

表 8 两种基于深度学习的生存对抗技术分析与对比

生存对抗技术	相同	不同
基于对抗样本生成的自动化免杀	1）目的相同，即用于恶意代码的生存； 2）都是通过深度学习模型分类功能来完成攻击。	1）强调对数据分布的变化； 2）要求攻击者具有对恶意代码操纵的能力。
深度学习伴生的后门攻击		1）强调对模型权重的修改； 2）攻击者拥有对触发后门的完全控制权。

其中最为核心的一点是：基于对抗样本生成的自动化免杀强调的是对数据分布的变化，而模型后门攻击的关注重点则在模型本身，强调对模型自身的改变（修改/激活某些神经元的权重）。除此之外，模型后门攻击能够为攻击者提供对后门触发器的完全控制权，因此后门攻击使得攻击者可以选择最方便的方式来触发错误分类，但对抗样本攻击则要求攻击者具有一定能力来操纵恶意代码以执行攻击[58]。

第二，深度学习模型本身存在一些“视觉盲点”，如图 10 所示，这些盲点位于模型决策边界两边的区域。由于模型训练中无法穷尽学习到所有可能的样本示例，因此，盲点区域的大小是未知的，这些未知的空间就很有可能被攻击者探索并利用。模型后门攻击很好地利用了这些盲点，通过修改模型中易受攻击的神经元的权重位，可以轻松建立后门触发器与指定神经元之间的强连接，从而导致模型的决策边界发生偏移，在保证对绝大部分正常样本做正确分类的前提下，又能够对携带后门触发器的恶意样本做出错误决策。如果盲点区域带给攻击者可探索的空间是巨大的，那么随着深度学习模型的广泛应用，及其具有可迁移应用的特性，模型后门攻击带来的安全风险也将是巨大的，因此针对模型后门攻击的安全防御研究当刻不容缓。

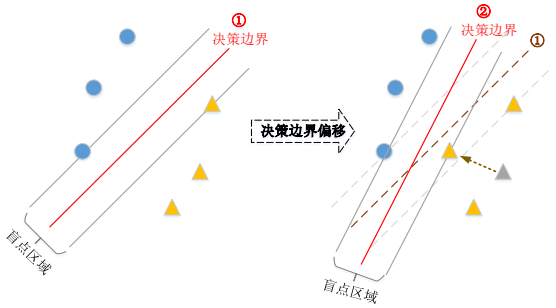

图 10 模型决策边界示意图

第三，模型后门攻击的风险不止表现在恶意代

码攻击链上，对于隐私数据的泄露问题，模型后门攻击带来的风险也值得引起关注，最近由 Tianyu Gu[59]等人通过假定恶意的外包服务提供商，对愈来愈热的神经网络外包培训带来的新的安全风险问题进行了探索，并证明了神经网络中后门功能的强大。但由于隐私泄露问题不是本文关注的问题，我们不做深入探究，只希望可以引起研究者的关注与重视。

4.2 深度学习伴生的模型后门防御

深度神经网络的可解释性和透明性的需求是当今计算的一大挑战。由于深度神经网络的黑盒性质引发的一个基本问题是无法详尽测试其行为。没有透明性，就无法保证模型在未经测试的输入上表现出预期的效果，这使得在深度神经网络中启用木马后门成为可能[46, 59]。由于神经网络木马后门功能的强大性，近年来，针对神经网络木马后门的有效防御措施也逐渐被提出。

2018 年，Kang Liu[60]等人将修剪和微调两种防御措施进行组合并提出 fine-pruning 的防御思路，并且还设计了一种可感知修剪的后门攻击对 fine-pruning 进行了评估，结果表明 fine-pruning 可以将攻击成功率降低到了 0%，首次实现了对 DNN 后门攻击的有效防御。同年，2018 年，Bryant Chen[58]等人提出了一种用于神经网络的后门检测和清除的新方法，该方法使用激活聚类来检测将后门插 DNN 模型的恶意样本，并且通过实验评估证明了激活聚类的方法在面向不同应用时的有效检测性能。

2019 年 Bolun Wang[61]等人他们通过详细的实验开发和验证了三种缓解方法：1）对抗输入的早期过滤器，用于识别具有已知触发器的输入，2）基于神经元修剪的模型修补算法，3）基于取消学习的模型修补算法，并且通过对各种 DNN 进行广泛的实验证明了这些防御措施的有效性。同年，Huili Chen[62]等人提出了第一个后门检测框架 DeepInspect，在没有干净的训练数据或真实参考模型的帮助下，检查预先训练的 DNN 的安全性，使用条件 GAN 学习潜在触发器的概率分布，通过扰动程度来检测后门插入痕迹，并且通过模型修补有效降低了触发器的激活率，实验评估表明，与以前的工作相比，DeepInspect 可提供卓越的检测性能和更低的运行时间开销。

综上，模型后门防御重点关注深度神经网络模型的健壮性问题。如表 9 所示，本文在对上述防御技术进行归纳总结后，发现当前的后门防御方法可以简单概括为两种：一种是数据清洗，即通过过滤触发后门的输入，降低触发后门的风险，典型的操作就是激活聚类的方法；另一种是模型修复，即在保证模型正常功能的前提下，对原有模型稍加改动，例如通过细粒度的修剪法，可以去除能够触发后门的神经元，以达到防御后门攻击的效果。

表 9 深度学习伴生的模型后门防御技术对比与分析

年份	文献	前提/假设	攻击引擎	数据集（来源）	攻击成功率	技术贡献
2018	[60]	白盒假设，假设强大的“白盒”攻击者可以完全控制训练过程和训练数据集。	DNN/CNN (DeepID 、AlexNet 、Faster-RCNN)	YouTube Aligned Face dataset、[46]、the U.S. traffic signs dataset	0%	提供了针对 DNN 的后门攻击的首个有效防御措施，迈出了防御深度神经网络中的后门攻击的第一步。
2018	[58]	1）白盒假设； 2）尽管后门样本和目标样本通过植入了后门的神经网络获得了相同的分类，但它们获得此分类的原因却有所不同，机制的差异在模型网络的激活层中应该很明显，这代表了网络如何做出决策。	DNN/CNN	MNIST 、LISA traffic sign Dataset 、Rotten Tomatoes movie reviews	接近 0%	在不需要任何可信数据的情况下检测模型后门攻击的第一种方法，并在不需要经过验证和信任数据集的情况下能够修复模型。

2019	[61]	1）假设大多数标签仍未受感染，即后门只能导致对少数输出标签(类)有针对性的错误分类； 2）白盒假设，假设防御者有权访问经过训练的 DNN，以及一组正确标记的样本，以测试模型的性能。防御者还可以使用计算资源来测试或修改 DNN。	DNN	MNIST、GTSRB、YouTube Face 、PubFig，VGGFace	<6.7%	提出了第一个针对 DNN 后门攻击的健壮且可推广的检测和缓解系统。
2019	[62]	1）白盒假设； 2）没有干净的训练数据或真实参考模型的帮助。	DNN (BadNets 、TrojanNN)	MNIST、GTSRB、VGGFace 、ResNet-18	0%	提出了第一个使用 GAN 学习潜在触发器概率分布的后门检测框架 DeepInspect。

虽然表 9 中的攻击成功率可以证明当前后门防御技术的有效性，但这些防御技术几乎都是基于已有的数据集，并且验证的是对已知攻击的防御能力，对于未知的模型后门攻击并不一定能有效地进行对抗。所以站在实际攻防对抗的角度，本文认为只要深度学习模型自身的安全问题一直存在，模型后门攻击就会有一直发生的可能，而对应地，模型后门防御将可能会一直处于一种被动防御的状态，并不能从根本实现对抗。因此本文从探索深度学习模型自身安全的角度出发，给出了几点思考与建议：

1）在未来深度学习的发展中，强调对深度学习模型算法的安全性是重要的，处理好深度学习泛化性与安全性之间的平衡问题，才会推动安全防御框架的发展。无论在模型开发的软件或硬件层面，编码漏洞或后门植入的可能性都是存在的，因此为了保证深度学习算法的安全性，还应当首先从底层基础架构的安全性上做到保障。

2）建议我国开发属于自己的人工智能基础架构和模型库。当前很多深度学习模型都是由美国科技巨头公开发布的，这些模型在全球范围内被广泛使用，来训练自己的算法，使得在人工智能领域，美国与其他国家拉开差距的速度比芯片领域更加迅速。国内现在很多公司依赖于美国的开放资源，当我们围绕着这些开源工具来形成自己的产品时，其结果就是为这些产品建立生态，为这些产品积累数据。一旦别人收走工具的使用许可，收走算法的使用权，那就再也来不及建立新的生态链、积累新的数据库了，这将导致我国在人工智能方面变得一无所有。所以从零实现并开发自己的深度学习模型的必要性与紧迫性是不言而喻的。

4.3 模型后门攻防研究的助力作用

如表 7 和表 9 中的“数据集（来源）”所示，当前模型后门攻击技术的数据对象多为文本、图片、音频、视频等，直接以恶意代码作为数据对象的后门攻防研究相对较少。但本文认为当前深度学习伴生的模型后门攻防研究对促进恶意代码的攻击和防御研究具有很好的助力作用，具体体现为：

1） 当前这些模型后门攻防技术看似与恶意代码的攻防对抗的相关性不是很大，但由于将恶意代码转换为图片或矩阵的形式进行分类检测的研究早在 2015 年 Kaggle 举办的 Microsoft Malware Classification Challenge (BIG 2015)[2] 大赛上就已经非常流行，这类恶意代码检测系统如果被部署在防御侧，现有的模型后门攻击技术则均可被攻击者利用，从而实现基于模型后门的自动化免杀。

2） 深度学习的黑盒特性使得模型后门成为一种几乎必然会衍生出的安全风险。而且由于深度学习的可迁移性和当前深度学习模型的公开可共享性的提升，接受各类恶意代码形式（比如 PE 文件、二进制文件、JS 脚本等）的不同恶意代码查杀模型可被植入后门也被认为是近乎必然的一件事情，从这个角度出发，攻击者也可以通过模型后门攻击助力实现恶意代码的自动化免杀。

3） 对于防御者而言，为了缓解模型后门攻击问题，他们会利用各类模型后门防御技术对部署的检测系统做修剪或微调等模型修补工作，修补后的模型将具有更少的“视觉盲点”，当这些模型被应用于恶意代码查杀、自动化网络钓鱼识别或者深度学习赋能的恶意行为检测时，可助力实现更具有健壮性防御系统。

5 结论与展望

当前深度学习赋能的恶意代码攻防研究已是网络空间安全领域中的热点问题[4, 7, 8, 27, 28, 44, 45, 63]，基于恶意代码攻击链定位深度学习在攻防研究中的赋能点，探索新型恶意代码的生成机理、入侵方式、攻击释放原理及生存对抗特性等是恶意代码攻防研究中的重点问题，而探索不同赋能攻击环节上的防御方法则是主要目标，这两部分的结合构成了深度学习赋能的恶意代码攻防研究中的主要研究方向。

为了及时跟进深度学习赋能的恶意代码攻防研究的最新进展，围绕恶意代码攻击链诠释的攻击过程，本文在可定位的赋能点上，对深度学习助力攻击、助力防御的新型技术展开了分析介绍；并且为了让深度学习发挥更佳的助力作用，本文也对因深度学习的自身脆弱性带来的伴生安全问题进行了关注。更重要地，本文从技术原理、实际可行性、面临的难题，以及未来发展趋势等多种不同的角度对上述技术展开了深入剖析，旨在为安全研究人员提供一些预测新型攻击手段和探索新型防御技术的启发。

除此之外，本文对深度学习在赋能恶意代码攻防研究中的发展趋势也有一些预见性的思考。尤其是在赋能攻击方面，深度学习技术将使恶意代码攻击变得更加复杂且难以检测。在以基于神经网络的精准定位与打击、基于生成对抗网络的流量模仿以及基于黑盒模型的攻击意图隐藏为代表的三种新型攻击技术被预测并进行了可行性攻击验证之后，针对这三种攻击技术的防御措施却尚未被研究者提出。这已经开始拉开了网络攻击与防御领域在赋能研究中的差距，证明攻击者在人工智能领域已经可以进行比较超前的探索，并且其未知的可探索空间可能会非常大。尤其是基于神经网络的精准定位与打击，以及基于黑盒模型的攻击意图隐藏这两种新型攻击技术通过主动利用深度学习模型作为恶意代码攻击的组件，开启了攻击者主动利用神经网络的第一步。

如果研究者更多地关注诸如认知计算和深度学习在僵尸网络中应用，便可以注意到僵尸网络领域存在着新的赋能研究的可能性。在这种情况下，新的僵尸网络特征可能是僵尸程序的自治[64]，研究学者甚至预测了未来可能的两种智能僵尸网络，分别称为 Hivenet 和 Swarmbot，这也是主动利用深度神经网络的一种潜在威胁。由此，主动利用深度学习式地赋能攻击将日渐成为一种新的趋势，但目前尚未发现针对这类赋能攻击的可行的防御手段，因此对于更有力的安全防御技术的研究仍迫在眉睫。

由此，本文分析得出结论：当前对于深度学习赋能的攻击技术，其发展趋势已经发生了范式转变，即由被动式利用深度学习以绕过防御引擎转变为主动利用深度学习模型作为攻击组件。然而，根据本文对深度学习伴生安全问题的分析，这些深度学习模型的不足依旧存在，从对抗的角度出发，防御者便可以顺势利用这些不足展开对应的防御研究，当前诸如模型逆向、模型萃取等一些新型技术已经在人工智能领域被提出，类似这样的技术如果能够扩展应用于网络空间安全防御领域，本文认为将能够大幅度弥合网络攻防研究的差距。

综上，当前深度学习赋能恶意代码攻防研究才刚刚起步，基于恶意代码攻击链的更多可能的赋能攻击与防御点有待研究者继续探索发掘。另外，深度学习助力恶意代码攻防的一大挑战是数据集的限制，如何建立有效、公开的数据集供研究者使用，这是一个非常值得思考和研究的问题。未来，笔者将投入更大的精力研究未知的攻击技术并预先构筑防御体系，推进相关的理论和实践问题的解决。

参 考 文 献

[1] Seok S, Kim H. Visualized malware classification based-on convolutional neural network. Journal of The Korea Institute of Information Security & Cryptology, 2016, 26(1): 197-208.

[2] Kaggle. Microsoft malware classification challenge (big 2015), https://www.Kaggle.Com/c/malware-classification/.

[3] Rusak G, Al-Dujaili A, O'Reilly U-M. Ast-based deep learning for detecting malicious powershell//Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, 2018: 2276-2278.

[4] Grosse K, Papernot N, Manoharan P, et al. Adversarial perturbations against deep neural networks for malware classification. arXiv preprint arXiv:1606.04435, 2016.

[5] Hu W, Tan Y. Generating adversarial malware examples for black-box attacks based on gan. arXiv preprint arXiv:1702.05983, 2017.

[6] Hu W, Tan Y. Black-box attacks against rnn based malware detection algorithms//Proceedings of the Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence. New

Orleans, United States, 2018: 245-251.

[7] Kolosnjaji B, Demontis A, Biggio B, et al. Adversarial malware binaries: Evading deep learning for malware detection in executables//Proceedings of the 2018 26th European Signal Processing Conference (EUSIPCO). Rome, Italy, 2018: 533-537.

[8] Raff E, Barker J, Sylvester J, et al. Malware detection by eating a whole exe//Proceedings of the Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence. New Orleans, United States, 2018: 268-276.

[9] Anderson H S, Kharkar A, Filar B, et al. Learning to evade static pe machine learning malware models via reinforcement learning. arXiv preprint arXiv:1801.08917, 2018.

[10] Trend micro incorporated research paper 2012, https://www.Trendmicro.De/cloud-content/us/pdfs/security-intelligence/white-papers/wp-spear-phishing-email-most-favored-apt-attack-bait.Pdf. 2012.

[11] Phishing defense guide 2017, https://www.Ciosummits.Com/phishme-phishing-defense-guide_2017.Pdf. 2017.

[12] 2019 data breach investigations report, https://enterprise.Verizon.Com/resources/reports/2019-data-breach-investigations-report.Pdf. 2019.

[13] 2013 年中国互联网网络安全报告, http://www.Cac.Gov.Cn/files/pdf/wlaq/annual%20report/2013%20annual%20report%20.Pdf. 2013.

[14] Baki S, Verma R, Mukherjee A, et al. Scaling and effectiveness of email masquerade attacks: Exploiting natural language generation//Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. Abu Dhabi, United Arab Emirates, 2017: 469-482.

[15] Das A, Verma R. Automated email generation for targeted attacks using natural language. arXiv preprint arXiv:1908.06893, 2019.

[16] Seymour J, Tully P. Weaponizing data science for social engineering: Automated e2e spear phishing on twitter//Proceedings of the Black Hat USA. Las Vegas, United States, 2016: 1-8.

[17] Seymour J, Tully P. Generative models for spear phishing posts on social media. arXiv preprint arXiv:1802.05196, 2018.

[18] antisnatchor. Practical phishing automation with phishlulz//Proceedings of the Kiwicon X. Wellington, New Zealand, 2016.

[19] Orru M, Trotta G. Muraena: The unexpected phish//Proceedings of the Hack In The Box Security Conference 2019. Amsterdam, Netherlands, 2019.

[20] Kirat D, Jang J, Stoecklin M P. Deeplocker - concealing targeted attacks with ai locksmithing//Proceedings of the Black Hat USA 2018. Las Vegas, United States, 2018: 1-29.

[21] Rigaki M, Garcia S. Bringing a gan to a knife-fight: Adapting malware communication to avoid detection//Proceedings of the 2018 IEEE Security and Privacy Workshops (SPW). San Francisco, United States, 2018: 70-75.

[22] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets//Advances in neural information processing systems, 2014: 2672-2680.

[23] Lin Z, Shi Y, Xue Z. Idsgan: Generative adversarial networks for attack generation against intrusion detection. arXiv preprint arXiv:1809.02077, 2018.

[24] Lee W, Noh B, Kim Y, et al. Generation of network traffic using wgan-gp and a dft filter for resolving data imbalance//Proceedings of the International Conference on Internet and Distributed Computing Systems. Naples, Italy, 2019: 306-317.

[25] Ring M S D, Landes D,, et al. Flow-based network traffic generation using generative adversarial networks. Computers & Security, 2019, 82: 156-172.

[26] Li J, Zhou L, Li H, et al. Dynamic traffic feature camouflaging via generative adversarial networks//Proceedings of the 2019 IEEE Conference on Communications and Network Security (CNS). Washington, D.C., United States, 2019: 268-276.

[27] 张思思, 左信, 刘建伟. 深度学习中的对抗样本问题. 计算机学报, 2019, 42(8): 1886-1904.

[28] Al-Dujaili A, Huang A, Hemberg E, et al. Adversarial deep learning for robust detection of binary encoded malware//Proceedings of the 2018 IEEE Security and Privacy Workshops (SPW). San Francisco, United States, 2018: 76-82.

[29] Xu W, Evans D, Qi Y. Feature squeezing: Detecting adversarial examples in deep neural networks. arXiv preprint arXiv:1704.01155, 2017.

[30] Wang Y, Stokes J W, Marinescu M. Neural malware control with deep reinforcement learning//Proceedings of the ICLR 2019. New Orleans, United States, 2019: 1-16.

[31] Jordaney R, Sharad K, Dash S K, et al. Transcend: Detecting concept drift in malware classification models//26th {USENIX} Security Symposium ({USENIX} Security 17), 2017: 625-642.

[32] Pendlebury F, Pierazzi F, Jordaney R, et al. {tesseract}: Eliminating experimental bias in malware classification across space and time//28th {USENIX} Security Symposium ({USENIX} Security 19), 2019: 729-746.

[33] Kudugunta S, Ferrara E. Deep neural networks for bot detection. Information Sciences, 2018, 467: 312-322.

[34] Soon G K, Chiang L C, On C K, et al. Comparison of ensemble simple feedforward neural network and deep learning neural network on phishing detection. Computational science and technology: Springer, 2020: 595-604.

[35] Abokhodair N, Yoo D, McDonald D W. Dissecting a social botnet: Growth, content and influence in twitter//Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing. vancouver, canada, 2015: 839-851.

[36] Echeverría J, Zhou S. The 'star wars' botnet with> 350k twitter bots. arXiv preprint arXiv:1701.02405, 2017.

[37] Anti-phishing working group, phishing activity trends report 3rd quarter 2019. 2019.

[38] Homayoun S, Ahmadzadeh M, Hashemi S, et al. Botshark: A deep learning approach for botnet traffic detection//Ali Dehghantanha M C, Tooska Dargahi. Cyber threat intelligence. Cham: Springer, 2018: 137-153.

[39] Marín G, Casas P, Capdehourat G. Deep in the dark-deep learning-based malware traffic detection without expert knowledge//Proceedings of the 2019 IEEE Security and Privacy Workshops (SPW). San Francisco, United States, 2019: 36-42.

[40] Prasse P, Machlica L, Pevný T, et al. Malware detection by analysing encrypted network traffic with neural networks//Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2017: 73-88.

[41] Lastline. Network security in an encrypted world, https://go.Lastline.Com/networksecurity_encryptedtraffic.Html?Utm_source=blog&utm_medium=whitepaper&campaign=netsec-encryptedtraffic. 2020.

[42] Tobiyama S, Yamaguchi Y, Shimada H, et al. Malware detection with deep neural network using process behavior//Proceedings of the 2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC). Atlanta, United States, 2016: 577-582.

[43] Rhode M, Burnap P, Jones K. Early-stage malware prediction using recurrent neural networks. computers & security, 2018, 77: 578-594.

[44] Rege M, Mbah R B K. Machine learning for cyber defense and attack//DATA ANALYTICS 2018 : The Seventh International Conference on Data Analytics. Athens, Greece, 2018: 1-6.

[45] Kubovič O, Košinár P, Jánošík J. Can artificial intelligence power future malware? l ńea]. Available: https://www.welivesecurity.com/wp-content/uploads/2018/08/Can_AI_Power_Future_Malware.pdf, 2018.

[46] Liu Y, Ma S, Aafer Y, et al. Trojaning attack on neural networks//Proceedings of the Network and Distributed Systems Security (NDSS) Symposium 2018. San Diego, United States 2017: 1-15.

[47] Ji Y, Zhang X, Wang T. Backdoor attacks against learning systems//Proceedings of the 2017 IEEE Conference on Communications and Network Security (CNS). Las Vegas, United States, 2017: 1-9.

[48] Chen X, Liu C, Li B, et al. Targeted backdoor attacks on deep learning systems using data poisoning. arXiv preprint arXiv:1712.05526, 2017.

[49] Liao C, Zhong H, Squicciarini A, et al. Backdoor embedding in convolutional neural network models via invisible perturbation. arXiv preprint arXiv:1808.10307, 2018.

[50] Rakin A S, He Z, Fan D. Tbt: Targeted neural network attack with bit trojan. arXiv preprint arXiv:1909.05193, 2019.

[51] Parkhi O M, Vedaldi A, Zisserman A. Deep face recognition. British Machine Vision Association, 2015: 1-12.

[52] Graves A, Jaitly N. Towards end-to-end speech recognition with recurrent neural networks//Proceedings of the International conference on machine learning 2014, 2014: 1764-1772.

[53] Eidinger E, Enbar R, Hassner T. Age and gender estimation of unfiltered faces. IEEE Transactions on Information Forensics and Security, 2014, 9(12): 2170-2179.

[54] Pang B, Lee L, Vaithyanathan S. Thumbs up?: Sentiment classification using machine learning techniques//Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, 2002: 79-86.

[55] Huang G B, Mattar M, Berg T, et al. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments//Proceedings of the Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition. Marseille, France, 2008: 1-11.

[56] Panayotov V, Chen G, Povey D, et al. Librispeech: An asr corpus based on public domain audio books//Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brisbane, Australia, 2015: 5206-5210.

[57] Li X, Roth D. Learning question classifiers//Proceedings of the 19th international conference on Computational linguistics-Volume 1. Stroudsburg, United States, 2002: 1-7.

[58] Chen B, Carvalho W, Baracaldo N, et al. Detecting backdoor attacks on deep neural networks by activation clustering. arXiv preprint arXiv:1811.03728, 2018.

[59] Gu T, Dolan-Gavitt B, Garg S. Badnets: Identifying vulnerabilities in the machine learning model supply chain. arXiv preprint arXiv:1708.06733, 2017.

[60] Liu K, Dolan-Gavitt B, Garg S. Fine-pruning: Defending against backdooring attacks on deep neural networks//Proceedings of the International Symposium on Research in Attacks, Intrusions, and Defenses. Heraklion, Greece, 2018: 273-294.

[61] Wang B, Yao Y, Shan S, et al. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks//Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP), 2019: 707-723.

[62] Chen H, Fu C, Zhao J, et al. Deepinspect: A black-box trojan detection and mitigation framework for deep neural networks//Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China, 2019: 4658-4664.

[63] Latah M. When deep learning meets security. arXiv preprint arXiv:1807.04739, 2018.

[64] Danziger M, Henriques M A A. Attacking and defending with intelligent botnets. XXXV Simpósio Brasileiro de Telecomunicaçoes e Processamento de Sinais-SBrT, 2017, 2017: 457-461.

JI Tian-Tian, born in 1995, Ph.D. candidate of cyberspace security with Beijing University of Posts and Telecommunications. Her current research interests include network security.

FANG Bin-Xing, born in 1960, Ph.D., Professor and Ph.D. supervisor. Academician of Chinese Academy of Engineering. His current research interests include computer architecture, computer network and information security.

CUI Xiang, born in 1978, Ph.D., Professor and Ph.D. supervisor. His main research interests include network security.

WANG Zhong-Ru, born in 1986, Ph.D. candidate of cyberspace security with Beijing University of Posts and Telecommunications. His current research interests include artificial intelligence and network security.

GAN Rui-Ling, born in 1996, M.S. candidate of cyberspace security with Beijing University of Posts and Telecommunications. Her current research interests include network security.

Han Yu, born in 1995, M.S. candidate of cyberspace security with Beijing University of Posts and Telecommunications. Her current research interests include network security.

YU Wei-Qiang, born in 1979, received the M.S. degree from Renmin University of China. His current research interests include network security, information security and artificial intelligence.

Background

With the rapid development of new technologies, such as Artificial Intelligence, Cloud Computing, Internet of Things and Big Data, cyberspace security is being confronted with a series of new threats and challenges. At present, the research on deep learning-powered malware attack and defense techniques has become a hot issue in the field of cyberspace security, and cybersecurity researchers have also made many significant research achievements in this field. However, for the application of deep learning in cyberspace security, there is no relevant review to follow up and summarize in time. Starting from the research on malware attack and defense, based on the analysis of the malware attack chain, this paper reviews, analyzes, and summarizes the application of deep learning in this field, aiming to promote the application of artificial intelligence in cyberspace security.

This work is supported by the Guangdong Provincial Natural Science Fund (Grant No. 2019B010136003), the Key Research and Development Program of Guangdong Province (Grant No. 2019B010137004), and the BUPT Excellent Ph.D. Students Foundation (Grant No. CX2019115).