基于双层注意力机制的恶意 URL 检测

赵云泽[1],蒋牧秋[2],董 伟[1],冯 志[1]
(1.华北计算机系统工程研究所,北京 100083;2.哈尔滨工业大学威海校区 经济管理学院,山东 威海 264209)

摘 要:随着信息化技术的不断发展,网络空间中存在的威胁也在不断变化。其中,基于恶意 URL 的攻击手段层出不穷。针对恶意 URL 识别与检测问题进行了深入探究,设计并实现了具有双层注意力机制的 Bi-LSTM 网络模型对恶意 URL 进行识别和检测,并将其命名为 A2Bi-LSTM。该模型分别在字符级别及单词级别对恶意 URL 中包含的可疑内容进行注意力权值的计算,进一步提升了恶意 URL 的识别精度。实验结果表明,A2Bi-LSTM 对恶意 URL 的识别准确率达到 97%,相较于传统检测模型有着更好的检测效果,能够有效应对此类攻击威胁,有助于网络空间安全体系的构建。

关键词:恶意 URL;注意力机制;网络安全;深度学习
中图分类号:TP393 **文献标识码**:A **DOI**:10.19358/j.issn.2097-1788.2023.02.001
引用格式:赵云泽,蒋牧秋,董伟,等. 基于双层注意力机制的恶意 URL 检测[J]. 网络安全与数据治理,2023,42(2):3-8.

Malicious URL detection based on double attention mechanism

Zhao Yunze[1], Jiang Muqiu[2], Dong Wei[1], Feng Zhi[1]
(1.National Computer System Engineering Research Institute of China, Beijing 100083, China;
2.School of Economics and Management, Harbin Institute of Technology, Weihai 264209, China)

Abstract: With the continuous development of information technology, threats in cyberspace are also changing. Among them, attacks based on malicious URLs keep intruding. In this paper, the problem of malicious URL identification and detection is deeply explored, and a Bi-LSTM network model with a two-layer attention mechanism is designed and implemented to identify and detect malicious URLs, which is named A2Bi-LSTM. This model calculates the attention weight value of suspicious content contained in malicious URLs at the character level and word level, which further improves the recognition accuracy of malicious URLs. The experimental results show that the identification accuracy of A2Bi-LSTM for malicious URLs reaches 97%, which is better than the traditional detection model, and it can effectively deal with such threats as well as help the construction of cyberspace security system.

Key words: malicious URLs; attention mechanism; cyber security; deep learning

www.pcachina.com

0 引言

随着信息化技术的急速发展和普及,当前社会各行业的信息化建设和数字化转型面临着严峻网络安全挑战。针对企业、政府、金融业等行业的网络攻击行为日益增多,尤其在高级持续性威胁(Advanced Persistent Threats, APT)逐步成为网络威胁主要方式的今天,传统的入侵检测系统、防火墙等防御手段已经难以抵御高等级、高隐蔽性、有组织的网络攻击行为。在 MITRE 公司所发布的对抗战术、技术和常识(Adversarial Tactics, Techniques and Common Knowledge, ATT&CK)网络攻击行为知识库[1]的阐述下,由恶意 URL 为基础的钓鱼攻击、水坑攻击、中间人攻击等手段成为实施 APT 攻击的初始必要手段,因此,针对恶意 URL 检测与识别的研究成为网络安全行业关注的重点。

统一资源定位器(Uniform Resource Locator, URL)是一种互联网标记形式,用以向互联网用户指明其资源位置和访问方式,是 Internet 地址中的标准资源。Webroot 在 2019 年发布的威胁报告指出,即便是在安全域名上也发现了超过 40%的恶意 URL,相较于

2018 年,网络钓鱼攻击事件的比例上升 36%,钓鱼网站的数量也增加了 220%。根据卡巴斯基安全公司 2020 年的数据统计得知,全球用户计算机遭受至少一次网络恶意攻击事件的比例为 10.08%,其中包含至少 1.72 亿条 URL 被网络安全设备标记为恶意 URL,数量极其庞大。可以看出,恶意 URL 已经成为网络犯罪的主要基石,因此,构建快速、精准、可泛化的恶意 URL 检测模型成为保障网络安全的关键点。

1 研究基础

在恶意 URL 检测与识别问题研究的初期,基于黑名单的 URL 检测与防御方法是一种简洁而有效的手段,Prakash 等人[2]通过建设访问黑名单对已掌握的恶意 URL、IP 地址或关联字等进行匹配过滤,并以近似匹配算法为基础,提出一个 PhishNet 模型,对列入黑名单的恶意 URL 有着稳定的识别效果,该模型还基于已知信息主动进行黑名单的扩展,具有一定的未知恶意 URL 检测能力,但随着黑名单数量的增加,其计算成本巨大。还有学者提出通过图形化的方法[3]对目前已知的公共黑名单数据进行识别和检测,也达到了一定效果。后续又有学者提出了以 TF-IDF 为基础的 Cantina 方法[4]、SpyProxy[5]、Cujo[6]等基于内容检测的恶意 URL 识别方法。但上述检测方法对于已知恶意 URL 的防范效果较佳,却难以应对庞大数量的未知恶意 URL 检测需求。

随着机器学习技术的发展,有学者将此类技术应用到了恶意 URL 检测的研究中来。Vanhoenshoven 等[7]通过多种不同的统计特征,使用决策树、朴素贝叶斯、KVM 以及多层感知机等机器学习算法,对恶意 URL 进行检测,在其实验数据集上,随机森林和多层感知机方法取得了最优效果。Azeez 等人[8]则以 URL 所包含的单词、语法、主机名等内容信息作为特征,使用朴素贝叶斯分类模型进行恶意 URL 的判定。Naik[9]则基于梯度增强分类器 XGBoost 对其设定的 IP 地址、URL 长度等共 18 维特征进行分类器的训练,也能够达到一定的识别效果。然而基于机器学习方法的检测模型严重依赖人工选取特征的合理性与准确性。在耗费精力和时间的同时,不能真正分析恶意 URL 的隐藏特征,因此机器学习方法的识别准确率难以得到有效提升。

深度学习技术的出现,解决了人工提取特征的弊端,实现了 URL 中所包含的高纬隐藏特征的自动提取。在部分代表性工作中,Saxe 等人[10]利用字符嵌入和卷积神经网络(CNN)等方法对恶意 URL 进行特征提取并识别。Afzal 等人[11]提出了一种名为 URLdeepDetect 的混合深度学习方法,分别使用了 LSTM 神经网络和 k-means 聚类模型对恶意 URL 进行检测。Bahnsen 等人[12]使用了多个模型进行恶意 URL 的检测,提出循环神经网络在处理该问题上优于传统机器学习方法。Shivangi 等人[13]在 ANN 神经网络和长短期记忆网络的基础上,设计并实现了一款浏览器插件,进行实时的恶意 URL 检测。Le 等人[14]在卷积神经网络的基础上构建了一种恶意 URL 检测网络 URLNet,从字符级别和字级别对 CNN 网络进行优化。URLNet 分别针对恶意 URL 中的字符以及单词进行了不同层次的词嵌入操作,并将两次嵌入得到的多维向量进行融合,这种方法兼顾了字符和单词的语义表示,有效提升了检测的效果,同时也在检测问题上提供了新的思路。

通过对既往工作的研究和分析可以看出,基于深度学习的检测方法相比机器学习方法取得了更为优异的识别准确率,但既往工作对于 URL 本身所包含的序列模式和空间结构特性没有取得深入的研究成果,过于关注 URL 中的字符语义信息却丢失了其中单词级别的语义信息。URLNet 模型在一定程度上解决了语义表示不全的问题,但其采用的 CNN 网络对于长序列表达的效果不佳。因此,本文在前人工作基础上,采用双向长短期记忆网络和双层注意力机制,分别对 URL 中的字符级和单词级进行词嵌入操作并重新分配注意力权重,在保留双层语义信息的基础上,增强模型对于长序列信息的记忆能力,并将 URL 中的重点字符或词汇进行权重的增加,提升模型的检测识别能力。

2 本文方法

本文所提出的 A2Bi-LSTM 检测模型由输入层、嵌入层、2 层 Bi-LSTM 网络、2 层 Attention 层以及最终的输出层组成。模型以 URL 字符串作为输入,经过特定分割后由嵌入层进行词向量的转化,继而经由第一层 Bi-LSTM 网络进行初始特征的提取,再由第一层注意力机制对字符向量进行权值分配,形成字符级别向量并作为第二层 Bi-LSTM 网络的输入。后续经由第二层 Bi-LSTM 网络后,输入第二层注意力机制,计算并得到单词级别的权重分配与特征向量,送入最终的输出层进行归一化处理,得到当前 URL 属于恶意 URL 的概率。模型结构如图 1 所示。

www.pcachina.com

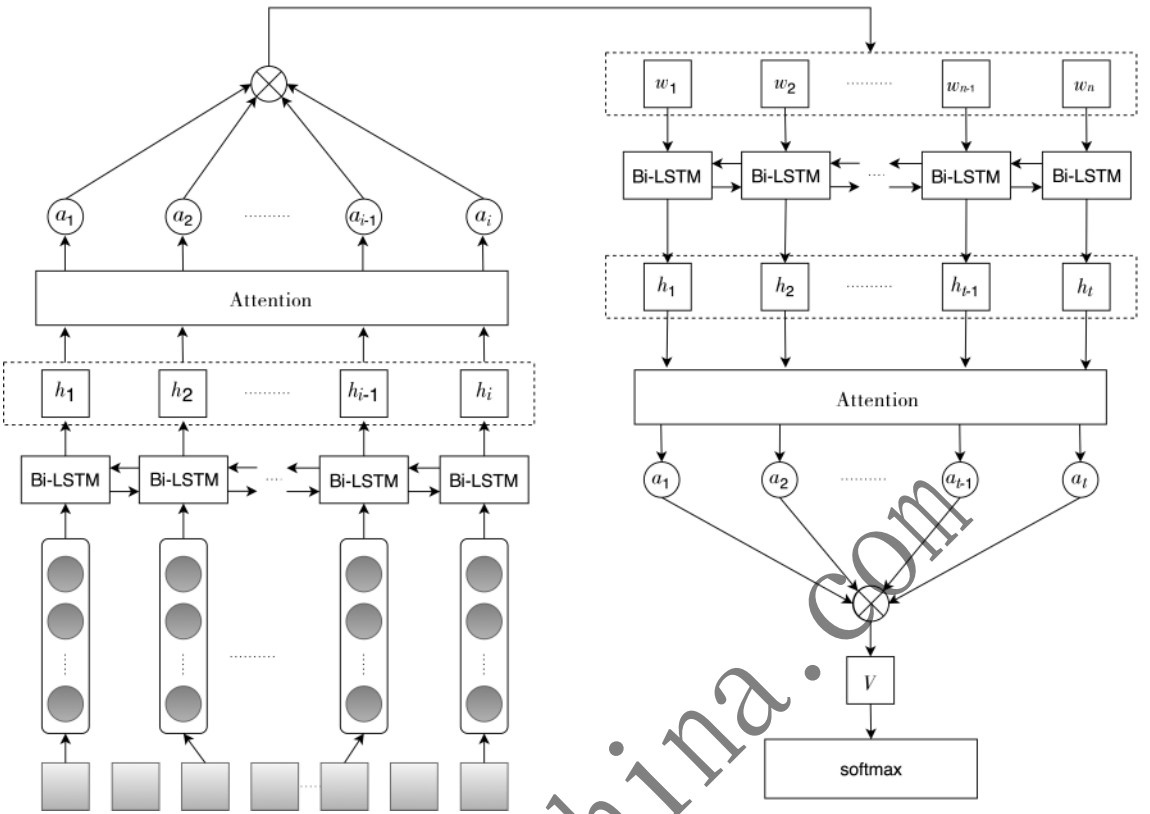

图 1　A2Bi－LSTM 模型结构图

2.1 输入层

对于一条 URL 来说，通常其组成部分包括协议、主机名、域名以及参数部分，如图 2 所示。

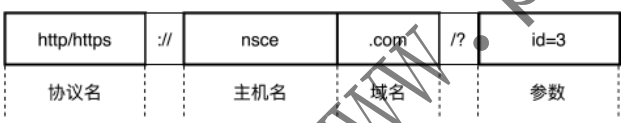

图 2　URL 结构图

其中，协议名位置通常使用 http 或 https 用以标明访问该 URL 所使用的协议类型；主机名与域名组合，构成其访问的主要位置；参数则表明了需要传递的参数。三者之间使用“/”进行分隔，参数之间常用的分割符除“？”之外，还包括“=”“&”等特殊符号。由于协议名部分与主机名、域名部分对于区分正常 URL 与恶意 URL 的区别影响较小，因此在本次实验的过程中，除去这三个部分，将识别的重点放置在参数部分。

根据 URL 的组成形式，将一组特殊符号定义为数组，其共包含出现在恶意 URL 中的常见的 13 种特殊分割符号，$S=\{'\&','\#','*','+','-','.','/',':','<','=','>','?','\_'\}$。依据 S 中的特殊符号对待输入的 URL 样本进行分割，可将一条 URL 表示为一条令牌，如式(1)所示：

$$T=\{t_1 \oplus t_2 \oplus \cdots \oplus t_n\} \tag{1}$$

2.2 嵌入层

word2vec 是谷歌于 2013 年开源的词向量特征训练工具[15]。它结合了神经网络语言模型和对数线性的优点，并利用分布式表示作为词向量的表示，该向量使用连续、密集的向量表示单词。

word2vec 提供了连续词袋(CBOW)模型和跳词(SKIP－GRAM)训练模型。CBOW 模型使用周围词用以预测中心词，而 SKIP－GRAM 模型则使用中心词预测周围词。由于 URL 不同位置的字符间可能存在一定的关联，CBOW 模型不能很好地提取字符间上下文含义，因此，本文使用了 SKIP－GRAM 模型作为词向量转换模型。

在针对大量恶意 URL 进行分析后发现，恶意 URL 的长度往往不超过 120，因此，在 Bi－LSTM 网络输入为等长序列的要求下，本文将输入的长度固定为 120，对高于此长度限制的 URL 删去多余的部

www.pcachina.com

分,不足此长度的 URL 则以“0”进行相应的补全,从而得到输入序列的词向量表示。

2.3 Bi-LSTM 层

URL 序列本身包含着相应的序列信息,为了避免传统循环神经网络在训练过程中导致的上下文语义丢失问题,本文在此处选择了双向长短期记忆网络(Bi-LSTM)对 URL 序列进行建模。

Bi-LSTM 实质上是一对 LSTM[16]的组合。其中前向 LSTM 从左到右执行序列以捕获即将到来的上下文,而向后 LSTM 从右到左执行顺序信息以捕获历史上下文。相较于单向 LSTM,双向 LSTM 网络能够对 URL 序列中的双向语义进行捕获,从而得到更为准确的语义信息。

Bi-LSTM 层的计算过程如式(2)~(4)所示:

$$\overrightarrow{h_t}=\overrightarrow{\mathrm{LSTM}}(V_n) \tag{2}$$

$$\overleftarrow{h_t}=\overleftarrow{\mathrm{LSTM}}(V_n) \tag{3}$$

$$h_t=[\overrightarrow{h_t},\overleftarrow{h_t}]^{\mathrm{T}} \tag{4}$$

2.4 注意力层

注意力机制(Attention)由 Vaswani 等人[17]首次提出。注意力机制模仿了人类大脑对于某一事件的注意力分配形式。通过设计特定的算法,将有限的注意力分配至应该关注的重点内容上,从而取得更佳的训练效果。

如图 3 所示,不难看出,在 NLP 等任务中,“man”为“i am a man”一句中的主要词汇。因此,当给定一个查询 Query(“man”),通过计算与 Key 的注意力分布得到“man”的注意力权重,并使得“man”一词在“i am a man”一句中的权重分配增加,从而获得对于任务训练中更为重要的数据信息进行输入,降低其他无关信息的权重占比,从而减小噪声干扰。

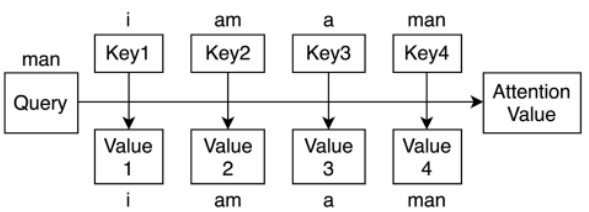

图 3 注意力机制计算示例

注意力权值的获取需要首先通过计算 Query 与每个 Key 的相似度,得到注意力权值矩阵,然后将注意力权重通过 softmax 函数进行[0,1]区间映射,继而根据归一化的注意力权重和 Value 得到最终的注意力权值。

字符级注意力层通过上述方法将 URL 中的关键字符进行注意力权值分配,从字符级别对输入双向 LSTM 模型的词向量计算权重分布,避免某些关键字符在训练过程中被模型忽略。

单词级注意力层则对第一层双向 LSTM 输出的组合后词向量矩阵进行权重的整体重新分配,进一步加大关键信息在训练中的权重占比。

2.5 输出层

输出层采用 softmax 函数对训练结果进行[0,1]区间映射,从而对输入的 URL 类别进行分类。当输出值大于固定阈值时,该 URL 被判定为恶意,当输出值小于固定阈值时,该 URL 被判定为良性。为防止模型过拟合,在输出层中加入了 Dropout 机制,以一定概率随机丢弃一些节点。

softmax 函数的计算方法如式(5)所示:

$$p=\mathrm{softmax}([wh_t+b]) \tag{5}$$

3 实验

3.1 数据集配置

为验证上述模型的有效性,本实验从 ISCX-URL-2016、Phishing Tank 以及 Kaggle 数据集所提供的数据集中合并整理了共 62 327 条样本数据。最终的用于实验的数据集内分为良性与恶意两类数据,其中良性样本 43 628 条,恶意样本 19 699 条,比例约为 7:3。设置训练集与测试集样本数量比例为 8:2。具体数据集分布如表 1 所示。

表 1 数据集详情

	训练集	测试集	总计
良性	34 902	8 726	43 628
恶意	15 759	3 940	19 699
总计	50 661	12 666	62 327

3.2 模型参数配置

实验使用 PyTorch 作为基本框架,PyTorch 作为目前主流的深度学习框架,为模型构建提供了一系列的便利支持。为提高模型训练速度,减少模型训练所需的时间消耗,本文采用了 Nvidia Tesla T4 GPU 用于此次模型的计算工作。通过多次训练,对模型所需的参数进行优化调整,得到最终的模型参数如表 2 所示。

3.3 实验设计

为验证本文所设计的模型在恶意 URL 识别问

www.pcachina.com

表 2 模型参数配置

超参数	大小
word2vec 词向量长度	120
Batch_Size 批量大小	128
LSTM 隐藏节点数	128
优化函数	Adam
Dropout 丢弃率	0.2

题上的效果提升，本文分别设计了基于支持向量机(SVM)的机器学习模型、基于单层 Bi-LSTM 的深度学习模型以及基于单层注意力机制的 Bi-LSTM 深度学习模型进行横向对比实验。其中 SVM 模型基于 TF-IDF 词频统计对 URL 特征进行提取，后续输入 SVM 模型进行分类；基于单层 Bi-LSTM 的深度学习模型使用 word2vec 进行词向量的生成，后续使用单层 Bi-LSTM 网络进行分类，其中不包含注意力层；基于单层注意力机制的 Bi-LSTM 深度学习模型与基于单层 Bi-LSTM 深度学习模型相比，仅增添一层注意力机制，其他保持一致。

实验均采用了 10 折交叉验证法对模型的性能进行评估，将原始数据集分为 10 份，依次选取其中的 9 份作为训练数据集，其余 1 份作为测试数据集。重复上述步骤 10 次，并取每次的模型训练结果的指标之和进行平均值计算，以此作为最终的评价指标值。

在本次实验中，分别使用 accuracy(准确率)、precision(精确度)、recall(召回率)和 $F1$ 分值组成的混淆矩阵作为分类评估指标对本文所设计的恶意 URL 检测模型的性能进行评估。

混淆矩阵中各指标的计算方法如式(6)~式(9)所示：

$$\text{accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}} \tag{6}$$

$$\text{precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{7}$$

$$\text{recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{8}$$

$$F1=\frac{2\times\text{precision}\times\text{recall}}{\text{precison}+\text{recall}} \tag{9}$$

其中 TP 为真阳性，TN 为真阴性，FN 为假阴性，FP 为假阳性。

4 实验结果与分析

通过上节所述的实验配置，实验得到 SVM、Bi-LSTM、单层注意力 Bi-LSTM 与本文所设计的 A2Bi-LSTM 在准确率、精密度、召回率和 $F1$ 值上的对比情况。

如图 4 所示，本文所设计的 A2Bi-LSTM 模型在此数据集上表现出的准确率比 SVM 模型高出 4.94%，比 Bi-LSTM 模型高出 4.1%，比单层注意力 Bi-LSTM 高出 1.14%；在精准率上，则分别高出SVM 模型、Bi-LSTM 模型 5.66%、5.35%；召回率分别高出 7.34%、5.10%、0.39%；$F1$ 值分别高出 7.34%、5.10%、0.39%。其中，在与 SVM 模型准确率的对比上，由于机器学习方法极其依赖特征的选择，在不当或错误特征的影响下，SVM 的有效特征提取不足，而深度神经网络模型可以将 URL 中的语义信息提取成高纬隐藏特征空间，获取更多有效特征，因此导致 SVM 模型的识别效果明显弱于 A2Bi-LSTM。与 Bi-LSTM 模型和单层注意力 Bi-LSTM 模型相比，A2Bi-LSTM 在各个指标上都有着更为优异的结果，单层注意力 Bi-LSTM 模型与 A2Bi-LSTM 模型结合了注意力机制和 Bi-LSTM 网络的优点，使得模型在训练过程中可以记忆 URL 序列中的长距离依赖信息，并通过注意力机制对序列中的信息进行更佳的权重分配，提升模型的识别效果。而 A2Bi-LSTM 模型与单层注意力 Bi-LSTM 模型相比，由于对权重进行了更为细致的分配，模型对于关键部分的注意力更加集中，因此也取得了更佳的识别效果。

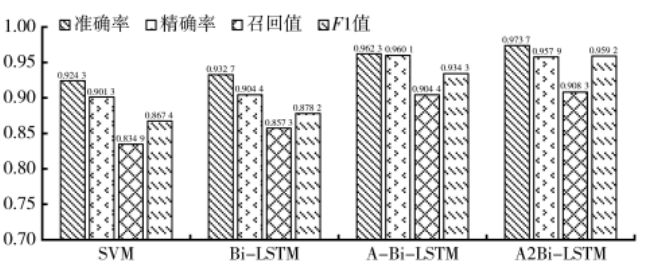

图 4 各模型评价指标对比

5 结论

本文提出了一种恶意 URL 检测模型 A2Bi-LSTM，该模型通过对恶意 URL 的层次语义特征进行提取，并在 Phishing Tank 等三个公开数据集所整合的数据集上进行了模型的训练与验证。实验结果表明，相较于传统恶意 URL 识别模型，本文所设计的检测模型能够有效地提升检测准确率，对于新型、变种的恶意 URL 数据仍能达到较好的识别效果，

www.pcachina.com

对后续的恶意URL识别的研究工作和应用具有一定的参考意义。

参考文献

[1] STROM B E, APPLEBAUM A, MILLER D P, et al. Mitre ATT&CK: design and philosophy[M]. Technical Report. The MITRE Corporation, 2018.

[2] PRAKASH P, KUMAR M, KOMPELLA R R, et al. Phishnet: predictive blacklisting to detect phishing attacks[C]//2010 Proceedings IEEE INFOCOM. IEEE, 2010: 1-5.

[3] KÜHRER M, ROSSOW C, HOLZ T. Paint it black: evaluating the effectiveness of malware blacklists[C]// International Workshop on Recent Advances in Intrusion Detection. Springer, Cham, 2014: 1-21.

[4] ZHANG Y, HONG J I, CRANOR L F. Cantina: a content-based approach to detecting phishing web sites[C]// Proceedings of the 16th International Conference on World Wide Web, 2007: 639-648.

[5] MOSHCHUK A, BRAGIN T, DEVILLE D, et al. SpyProxy: execution-based detection of malicious web content[C]// USENIX Security Symposium, 2007: 1-16.

[6] LIANG C, LIN H, WANG Q, et al. A redox-active covalent organic framework for the effcient detection and removal of hydrazine[J]. Journal of Hazardous Materials, 2020(381): 120983.

[7] VANHOENSHOVEN F, NÁPOLES G, FALCON R, et al. Detecting malicious URLs using machine learning techniques[C]//2016 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, 2016: 1-8.

[8] AZEEZ N A, SALAUDEEN B B, MISRA S, et al. Identifying phishing attacks in communication networks using URL consistency features[J]. International Journal of Electronic Security and Digital Forensics, 2020, 12(2): 200-213.

[9] NAIK N N. Modelling enhanced phishing detection using XGBoost[D]. Dublin: National College of Ireland, 2021.

[10] SAXE J, BERLIN K. eXpose: a character-level convolutional neural network with embeddings for detecting malicious URLs, file paths and registry keys[J]. arXiv preprint arXiv: 1702.08568, 2017.

[11] AFZAL S, ASIM M, JAVED A R, et al. URLdeepDetect: a deep learning approach for detecting malicious URLs using semantic vector models[J]. Netw. Syst. Manage., 2021, 29(21).

[12] BAHNSEN A C, BOHORQUEZ E C, VILLEGAS S, et al. Classifying phishing URLs using recurrent neural networks[C]//2017 APWG Symposium on Electronic Crime Research (eCrime). IEEE, 2017: 1-8.

[13] SHIVANGI S, DEBNATH P, SAJEEVAN K, et al. Chrome extension for malicious URLs detection in social media applications using artificial neural networks and long short term memory networks[C]//2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE, 2018: 1993-1997.

[14] LE H, PHAM Q, SAHOO D, et al. URLNet: learning a URL representation with deep learning for malicious URL detection[J]. arXiv preprint arXiv: 1802.03162, 2018.

[15] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[J]. Advances in Neural Information Processing Systems, arXiv preprint arXiv: 1310.4546, 2013.

[16] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[17] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, arXiv preprint arXiv: 1706.03762, 2017.

(收稿日期:2022-10-31)

作者简介:

赵云泽(1995-),男,硕士,助理工程师,主要研究方向:网络安全。

蒋牧秋(2000-),男,学士,主要研究方向:信息处理。

董伟(1986-),通信作者,男,硕士,高级工程师,主要研究方向:网络安全、工控安全。E-mail:dongwei@ncse.com.cn。

www.pcachina.com

版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部

www.pcachina.com