计算机科学
COMPUTER SCIENCE

DUWe:动态未知词嵌入方法在Web异常检测中的应用

王丽，陈刚，夏明山，胡皓

引用本文

王丽，陈刚，夏明山，胡皓. DUWe:动态未知词嵌入方法在Web异常检测中的应用[J]. 计算机科学, 2024, 51(6A): 230300191-5.

WANG Li, CHEN Gang, XIA Mingshan, HU Hao. DUWe:Dynamic Unknown Word Embedding Approach for Web Anomaly Detection [J]. Computer Science, 2024, 51(6A): 230300191-5.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230300191

DUWe:动态未知词嵌入方法在 Web 异常检测中的应用

王　丽[1,2,3]　陈　刚[1,3]　夏明山[1,2]　胡　皓[1]
1 中国科学院高能物理研究所　北京 100049
2 散裂中子源科学中心　广东 东莞 523803
3 中国科学院大学　北京 100049

摘　要　现有的基于深度学习模型的词嵌入方法用于 Web 异常检测时,通常将语料库中没有出现的未知词汇(Out of Vocabulary,OOV)设置为 unknown,并赋予零或随机向量输入到模型中进行训练,未考虑未知词汇在 Web 请求语句中的上下文关系。同时,在 Web 系统代码开发过程中,基于个人习惯并为了增加代码的可读性,程序员设计的请求路径代码往往存在一定的模式。因此,考虑到 Web 请求的模式和单词语义间的相关性,研究基于 Word2vec 的动态未知词表示方法 DUWe(Dynamic Unknown Word Embedding),该方法通过分析 Web 请求路径中单词上下文的关系来赋予未知词向量的表示内容。在 CSIC-2010 和 WAF Dataset 数据集上的实验评估表明,增加未知词表示方法比仅用 Word2vec 静态特征提取方法具有更好的性能,同时在准确性、精准率、召回率和 F1-Score 方面均有提高,在训练时间上最大降低 1.14 倍。

关键词:未知词汇;Web 异常检测;动态词嵌入;词嵌入优化;深度学习
中图分类号　TP393

DUWe:Dynamic Unknown Word Embedding Approach for Web Anomaly Detection

WANG Li[1,2,3],CHEN Gang[1,3],XIA Mingshan[1,2] and HU Hao[1]
1 Institute of High Energy Physics,Chinese Academy of Sciences(CAS),Beijing 100049,China
2 Spallation Neutron Source Science Center(SNSSC),Dongguan,Guangzhou 523803,China
3 University of Chinese Academy of Sciences,Beijing 100049,China

Abstract　When the existing deep-learning model-based word embedding methods are used to detect Web anomalies,the vocabulary not appearing in the corpus is usually called out of vocabulary(OOV) and is set as unknown,and given zero or random vector as the input of the depth model for training without considering the context of unknown word in the web request. In the process of code development,in order to increase the readability of code,programmers often design request path code based on a certain pattern which usually makes web requests semantically related. Considering that there are certain request patterns in web requests and pattern correlation between semantics,this paper studies and proposes a dynamic unknown word embedding method DUWe based on Word2vec,which assigns unknown word representation through word context inference. Evaluation on CSIC-2010 and WAF dataset shows that adding unknown word embedding methods have better performance than word2vec feature extraction methods. The accuracy,precision,recall rate and F1-Score are improved,and the maximum reduction in training time is 1.14 times.

Keywords　Unknown word,Web anomaly detection,Dynamic unknown word embedding,Word embedding optimization,Deep learning

1　引言

Web 应用给人们生活带来了极大的便利性,如远程办公、视频会议、在线购物和娱乐等。然而其本身具有开放性、易访问性,很容易受到黑客的攻击,因此 Web 安全一直是网络安全研究的热门话题。OWASP 报告[1]指出,Web 攻击类型(如越权访问、SQL 注入攻击、跨站脚本攻击、WebShell 攻击等等)层出不穷,给 Web 安全防御带来了非常严峻的挑战。

为了预防和阻止 Web 攻击,研究人员对 Web 攻击的特点和防御技术进行了广泛深入的研究,并取得了很多成果,如早期的 Web 应用防火墙(WAF)[2]、入侵检测系统[3],这些技术一般基于一定的规则和专家知识进行检测,主要缺点是无法检测未知攻击。异常检测是基于网络攻击异常行为与正常行为可以区分的一种假设;但由于行为模型不容易被建立,导致异常检测方法通常会存在一定的误报率。

近年来,科学家更倾向于研究各种机器学习、深度学习和迁移学习等异常检测方法,用以保护 Web 应用程序免于攻击。最著名的用于 Web 安全的神经网络是长短期记忆循环神经网络(LSTM-RNN)和卷积神经网络(CNN)。为了达到更高的攻击检测准确率,神经网络输入数据必须对数据请求

基金项目:国家自然科学基金(11905239,12005248,12105303)
This work was supported by the National Natural Science Foundation of China(11905239,12005248,12105303).
通信作者:王丽(wangli320@ihep.ac.cn)

进行预处理。文献中最常用的处理 Web URL 请求的技术是单词嵌入[4]和字符嵌入[5]，这些都是基于自然语言处理的表征方法。虽然这些方法已经取得了很好的性能，但它们将 HTTP URL 请求视为由单词或字母组成的一般句子或者字符串，对每个单词或者字符进行平均关注；并且在处理未知词或未知字符的时候通常设置为 unkown，赋予其零向量或者随机向量，而忽略 URL 请求具有一定的逻辑意义，而不是由无意义的单词或者字符串组成。

本文提出了一种针对上述问题的解决方法——动态未知词嵌入模型 DUWe(Dynamic Unknown Word Embedding)。该方法是词嵌入模型的一种改进，首先通过对 Web 请求每条 URL 进行预处理，包括解码、基于特殊字符的分词处理；而后通过 Word2vec 进行训练，构建语料库得到每个单词的静态表示；利用语料库及 URL 请求中单词上下文及逻辑关系推断赋予未知词表示。这项工作的目标是通过对未知单词的表示研究来提高检测精度。

本文第 2 章深入介绍了 Web 攻击异常检测相关工作和挑战；第 3 章介绍了与本文研究内容相关的知识背景；第 4 章描述了本文研究方法的框架；第 5 章介绍了数据集、评估和比较结果；最后总结全文。

2 相关工作

Web 异常检测受到研究人员的广泛关注。在 2003－2006 年间，Kruegel 等[6-8]提出多种异常检测模型，利用 Web 请求样本进行参数的数值长度、特征分布、是否缺失、顺序、访问频率等多个统计特征，有效地识别出异常的网络行为。Tekerek 等[9]构造了一个复合 Web 应用程序防火墙(Web Application Firewall，WAF)，该 WAF 结合了基于异常识别与基于签名识别的方法，以提高 Web 异常检测的精度。Applebaum 等[10]对开源 WAF、基于签名的检测 WAF、基于异常检测的 WAF 以及混合检测 WAF 已发表工作进行总结和分析，从另一个角度揭示 Web 攻击检测的发展。Gao[11]，Suneetha[12]，Feng[13]等分别基于 Web 日志对用户异常行为展开研究，利用聚类算法、数据挖掘算法、SVM 机器学习算法等挖掘可疑的用户，以达到异常检测的目的。然而传统的机器学习算法受限于特征提取，为了获得更好的检测效果，通常需要花费大量的时间来研究和调整特征的选择。

随着深度学习方法在图像处理、语音识别、自然语言处理等领域取得显著成果，不少研究者[14-26]将深度学习模型应用到 Web 异常检测，他们通过实验证明了深度学习算法在检测 Web 应用程序攻击方面能达到很好的效果。Liang 等[14]提出了一种基于 RNN 的异常检测方法。受益于同时使用 CNN 和 LSTM，Jemal 等[15]提出了 M-CNN 模型，将 LSTM 巧妙地插入到 CNN 的层中，有利于 LSTM 存储单元、卷积和 CNN 的最大池化操作。文献[16]中同时对 URL 进行字符和单词表示，然后将两种表示进行了连接，最后采用 CNN 进行异常检测。Mickiaki 等[17]研究了字符级卷积神经网络(CLCNN)来检测 Web 应用程序攻击。每个字符都由 8 位数字表示，每条 HTPP 请求根据其所包含的字符进行表示，形成 128 维向量。Seyyar 等[18]针对 URL 检测提出了 BERT 模型用于语义分析阶段，使用 CNN 进行最终分类。Biodoumoye 等[19]建立 Web 入侵检测系统，使用 Distil-BERT，RNN 和 LSTM 模型，分别识别 body 攻击、URL 攻击和 User-Data 攻击，实验结果显示其能够识别多种混合攻击。文献[20]针对 Web 攻击检测深度学习模型，提出了一种保护隐私的分布式训练过程。该模型允许参与者共享训练过程，以提高深度模型用于 Web 攻击检测的准确性，同时保留本地数据和本地模型参数的隐私性。

为了达到更好的效果，研究人员专注于不同的特征选择方法，其中文献[15-17]采用了 ASCII 码嵌入(ASCII Embedding)方式，文献[14]采用了词嵌入模型(Word Embedding)方式，文献[18]在语义分析方面采用了 BERT 模型，而文献[21-23]研究了字符嵌入(Character Embeddings)方式，文献[24-25]采用了 n-gram 方式。因为深度学习具有工程特征自动提取和选择的特点，这也导致了特征提取存在一定的缺陷。本文就大部分文章中对未知词设定为 unkown 及零向量这一问题进行了研究。

3 知识背景

3.1 Web 流量日志

Web 日志文件如图 1 所示，其中记录了 Web 用户向 Web 服务器提交请求时的活动信息，所有的信息都和 URLs 这串请求内容相关。URL 通常由请求域名、请求路径和查询参数组成，其请求结构可表示为：http://＋hostname＋:＋port＋/＋req_path＋?＋req_query。

```
get http://localhost:8080/tienda1/publico/anadir.jsp?id=2&nombre=Jam
+ibrico&precio=85&cantidad='; +drop+table+usuarios; +select+ * +
from+datos+where+nombre+like+'%&b1=aadir+al+carrito
```

图 1 Web 请求

Fig. 1 Web request

在实际 URL 请求中，hostname 用于定义服务域名，req_path 用于定位资源，req_query 用于向资源传递参数。在异常请求中，hostname 和 port 很容易被发现，因为这两个内容必须是实际存在的，如果不存在，Web server 会返回 404 或者无法访问等。而在 req_path 中包含的所有路径都具有一定的相关性，因为程序员在代码开发中为了代码可读性会进行设置，如 CateController. java 中的 list()在请求的路径表示为 cate/list；而 req_query 也不是随机定义的，通常也有一定的词位置模式。这是一种开发习惯，目前大多数开发人员均保留这种习惯，以便加强代码的可读性，这也是我们对未知词定义的想法的来源。

3.2 嵌入(Embedding)模型

嵌入模型是自然语义处理核心方法之一，它能够将文本转换成数学向量，起到人与计算机之间进行语言分析的桥梁作用，在文本分类等方面发挥着重要作用。在 Web 异常检测中，嵌入模型表征对深度模型 Web 异常检测起到了关键的作用，如 Word embedding，Character Embeddings，ASCII embedding，它们从不同角度将单词、字符或者 ASCII 映射为数学向量用于深度模型的输入。

比较流行的 Word embedding，可以将单词映射为实数向量，使语义相似的词具有相似的向量表示，如使用 Word2vec 工具。紧随 Word2vec 之后，出现了一些改进和变种，如全局向量词嵌入(GloVe)和子词嵌入(FastText)。Word2vec 是一个更经典的语言模型，同样是一个神经网络模型，训练神经网

络得到网络的权值矩阵，可以快速有效地训练词向量。如图 2 所示，Word2vec 主要有两种形式：连续词袋模型（Continuous Bag of Words，CBOW[26]）和跳字模型（Skip-gram[27]）。其中 CBOW 是通过上下文来预测当前位置词，Skip-gram 则是通过当前词来预测上下文。本文中也是首先使用 Word2vec 模型之一的 CBOW 来用向量表示 URL 中的词，只是本文的重点在于对未知词的处理方面。

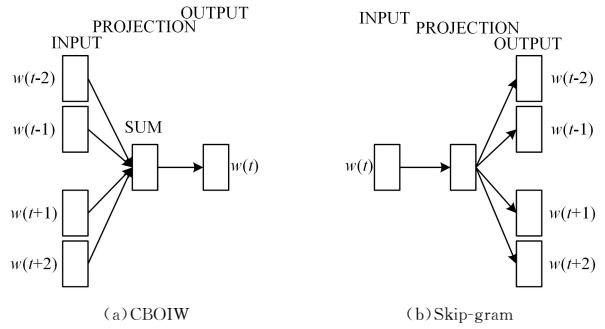

图 2　Word2vec 模型

Fig. 2　Word2vec model

4　研究方法

Web 异常检测框架主要由数据预处理模块、词嵌入学习模块和检测模块组成，如图 3 所示。

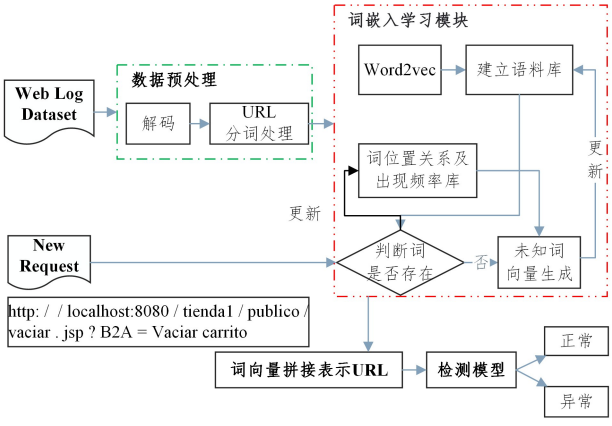

图 3　异常检测框架

Fig. 3　Framework of anomaly detection

4.1　数据预处理模块

在数据集 HTTP CSIC 2010[28] 中，我们主要提取 GET，PUT，POST 的请求。预处理模块是利用/，@，$，〈，〉，(，)，{，}，+，?，&，“，‘，～，\和；等特殊字符，将每个 URL 请求分割成一个单词串（包括一些特殊字符的字符串）。如果 URL 被编码，则需要先进行解码，而后再进行分词。

4.2　词嵌入学习模块

词嵌入学习模块主要是学习请求语料库中词的向量表示。在词嵌入学习模块，主要进行 3 方面的工作：单词向量及语料库的建立、词位置关系及两个词相邻出现的频率和未知词生成。

1）建立单词向量及语料库

将已经被分割成词的 Web 流量日志以一条条 URL 的形式传入 Word2vec 进行训练，得到单词向量及语料库记为：

$$W\langle w,v\rangle=\{\langle w_1,v_1\rangle,\langle w_2,v_2\rangle,\cdots,\langle w_n,v_n\rangle\}$$

其中，w_i 表示第 i 个词语，v_i 表示 w_i 映射的词向量。每个已知单词的向量表示：

$$v_i=f_{\text{word2vec}}(w_i) \tag{1}$$

2）词位置关系及出现频率

词位置关系是指两个单词组合一起出现的顺序位置关系。具体地，定义两个词 X，Y 的位置关系可表示为 XY，亦可表示为 YX，我们称之为二元词关系，并且认为这是两种位置关系。在这里，计算任意两个词 X，Y 组成二元词关系 XY 出现的频率可表示为：

$$TF_{XY}=\frac{t(wordXY)}{s} \tag{2}$$

其中，$t(wordXY)$ 表示该二元单词组在文件中出现的次数，s 表示所有二元单词组在文件中出现的次数之和。同时，根据请求中的词语动态更新二元词关系及出现频率。

3）DUWe：未知词生成向量目标函数

在很多论文中，未知词的特征表示比较简单，要么是零向量，要么是随机向量。在本文中，未知词的表示是研究重点。有以下几点说明：

1）具体的未知单词 Z 表示为 w_z，单词 X 和 Y 为已知，且单词 Z 和二元词关系 XY 前置相邻即 XYZ，我们认为 X 和 Y 对于 Z 的出现是均衡的。

2）未知词 Z 生成向量首先由其前置词关系表示，特别是对于 URL 来说，以 http://开始，处理后为 http:，/，/所以未知词 Z 前面一定会存在一个已知的前置二元词关系。尽管 XYZ，XZY 和 ZXY 这些词序列各自可能表达着不同的语义，但在确定了它们出现的顺序之后，未知词 Z 的生成向量主要是由 XY 序列决定的。这表明，即使是相同的单词组成，不同的排列顺序也会导致未知词 Z 的向量表示出现差异。。因此不同语义对分类结果有影响。

3）当出现 XYZ 及 XYZ'，而 Z 不等于 Z' 时，Z 和 Z' 的前置词关系都是 XY，无论 XYZ 及 XYZ' 哪一种先出现，XY 二元词关系统计量都会变化，即 TF 会动态更新，所以 Z 和 Z' 不会生成同样的词向量。

因此未知词 w_z 的二元词关系表示为：

$$g_{\text{wordReal}}(w_z)=TF_{XY}*\frac{1}{2}(v_X+v_Y) \tag{3}$$

综合式(1)－式(3)，单词向量表示为：

$$y(w_i)=\begin{cases}f_{\text{word2vec}}(w_i), & w_i \text{ in } W\\ g_{\text{wordReal}}(w_i), & w_i \text{ not in } W\end{cases} \tag{4}$$

5　实验与评价

我们在公开的标准数据集上进行了实验来评估所提方法的性能。实验环境是 Win10 操作系统，具有 i7 CPU 处理和 32 GB RAM，NVIDIA Quadro P600。

5.1　词嵌入学习模块

实验基于 2 个数据集：HTTP DATASET CSIC 2010[28] 和 WAF Dataset[29]。

1）HTTP DATASET CSIC 2010 被广泛用于 Web 入侵检测的研究。该数据集是 CSIC（西班牙研究全国委员会）在一个 Web 应用程序上自动收集的，包含 36 000 个正常请求和 25 000 多个恶意请求。HTTP 请求被标记为正常或异常，且包括 SQL 注入、缓冲区溢出、文件泄露、CRLF 注入、跨站脚本、参数篡改等各种攻击类型。对于原始的 HTTP 请求数据，主要提取出 GET，POST，PUT 请求数据用于检测，数据

集的 90%用于训练，剩下的用于测试。

2）WAF Dataset 作为一个标记的数据集，是 Github 上的一个开源项目，由 Faizan Ahmad 美国弗吉尼亚大学学生开源，机器学习驱动的 Web 应用防火墙研究，已经被广泛应用到很多学者的论文中。实验过程中，随机选择 45 101 条正常请求和 25 000 条恶意请求样本构造数据集，同样的，数据集的 90%用于训练，剩下的用于测试。

5.2 词嵌入学习模块

通过准确性、精准率、召回率和 F1-Score 评估结果，这是评估机器学习或深度学习中使用的模型的典型性能指标。准确率（Accuracy）表示所有请求被正确检测到的占所有数据的比例。召回率（Recall）是检测到异常攻击的真实攻击与所有攻击的比率。精准率（Precision）是检测到攻击的正常请求在所有正常请求中的占比。F1-Score 是用来衡量二分类模型精确度的一种指标，计算公式分别为式（5）—式（8）。

$$Accuracy=\frac{TP+TN}{TP+FP+TN+FN} \tag{5}$$

$$Precision=\frac{TP}{TP+FP} \tag{6}$$

$$Recall=\frac{TP}{TP+FN} \tag{7}$$

$$F1=\frac{2*(Precision*Recall)}{(Precision+Recall)} \tag{8}$$

5.3 实验结果和性能

为了验证所提出方法的有效性，本文选取了比较常见的深度学习模型进行对比分析。本文对 Word-CNN-GRU[18] 和 Word-CNN[19] 的深度学习算法进行了复现，对比了 Word-CNN-GRU 和 Word-CNN 增加 DUWe 前后的实验结果。除此之外，实现了 Word-RNN 模型，并对比了增加 DUWe 的实验结果。

需要说明，在上述样本集中所选用正常样本与异常样本数量相差较大。在训练过程中，我们将正样本与负样本混合作为一个整体，随机抽取该整体的 90%用于训练。10%用于测试，由于随机性较大，所选择的样本对实验结果有一定的影响。因此，实验过程中进行反复训练，每次训练迭代 10 次，经过多次运行得到最优结果。

从表 1 和表 2 的实验结果可以看出，DUWe 方法在结合深度学习模型进行异常检测时，对比用 Word2vec 静态特征提取、未知词设置为零向量的方法，检测效果提升显著，这说明增加 DUWe 的词嵌入模型能够从 URL 请求中捕获更多的语义信息，并且通过已知词进行未知词表征方法非常有效。特别是采用 RNN-DUWe 方法，在 CSIC2010 和 WAF 数据集上，对比其他深度学习模型，Accuracy，Precision，Recall，F1-Score 均呈现最优结果。这是由于 RNN 的优势是具有记忆功能，可以根据上一时刻的输入来影响这一时刻的输出，从侧面也证实了 URL 语句本身具有一定的请求模式，语句包含的词语之间具有一定的关联性。

表 1 深度学习模型增加 DUWe 前后在 CSIC2010 的实验结果

Table 1 Experimental results of deep learning model on CSIC2010 before and after adding DUWe

（%）

Algorithm	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Word-CNN-GRU	97.99	97.45	99.12	98.28
Word-CNN-GRU-DUWe	**99.08**	**99.18**	**99.24**	**99.22**
Word-CNN	93.40	91.84	98.59	95.10
Word-CNN-DUWe	**94.22**	**92.04**	**98.85**	**95.32**
RNN	96.79	96.26	98.28	97.26
RNN-DUWe	**99.38**	**99.28**	**99.67**	**99.47**

表 2 深度学习模型增加 DUWe 前后在 WAF 的实验结果

Table 2 Experimental results of deep learning model on WAF before and after adding DUWe

（%）

Algorithm	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Word-CNN-GRU	98.25	98.02	99.30	98.66
Word-CNN-GRU-DUWe	**98.30**	**98.04**	**99.36**	**98.70**
Word-CNN	97.46	97.16	98.91	98.03
Word-CNN-DUWe	**97.83**	**97.30**	**99.35**	**98.31**
RNN	98.49	98.08	99.63	98.85
RNN-DUWe	**99.64**	**99.57**	**99.87**	**99.72**

增加 DUWe 前后的训练时间结果如表 3 和表 4 所列。可以看出，增加 DUWe 后所有深度模型的训练时间均减少，且 Word-CNN-GRU-DUWe 对比 Word-CNN-GRU 下降最多，约为 1.14 倍，这是由于在训练过程中未知词根据已知词关系建立了词向量，使得该未知词变成了已知词，减少了遍历时间。

表 3 WAF 训练时间结果

Table 3 WAF training time

Algorithm	Word-CNN-GRU	Word-CNN-GRU-DUWe	RNN	RNN-DUWe	Word-CNN	Word-CNN-DUWe
训练时间（迭代 10 次）/s	716.59	**334.75**	310.90	**197.10**	909.50	**708.02**

表 4 CSIC2010 训练时间结果

Table 4 CSIC2010 training time

Algorithm	Word-CNN-GRU	Word-CNN-GRU-DUWe	RNN	RNN-DUWe	Word-CNN	Word-CNN-DUWe
训练时间（迭代 10 次）/s	532.62	**523.52**	389.80	**386.28**	636.59	**624.93**

结束语 本文中提出了一种改进的单词嵌入学习模型来自动学习请求的向量表示，不仅不需要任何专家知识和手工设计的特征，还弥补了未知词向量生成的缺陷，通过词语关系自动生成未知词向量表示。在不同数据集及不同深度学习模型上进行了增加 DUWe 前后的异常检测效果验证，实验结果表明，该方法在 HTTP CSIC 2010 和 WAF Dataset 数据集上的检测效果显著，优于现有的检测方法；除此之外，对增加 DUWe 前后的训练时间进行了对比，结果表明增加 DUWe 大大节省了训练时间。下一步将在真实的数据集上开展研究工作。

参 考 文 献

[1] The Open Web Application Security Project. OWASP Top 10: 2021, [online] Available: https://owasp.org/www-project-top-ten/.

[2] PROKHORENKO V, CHOO K K R, ASHMAN H. Web appli-

cation protection techniques：a taxonomy[J]. J. Netw. Comput. Appl.，2016，60：95-112.

[3] KUMAR K N，SUKUMARAN S. A survey on network intrusion detection system techniques[J]. Int. J. Adv. Technol. Eng. Explor.，2018，5(47)：385-393.

[4] LEBRET R P. Word embeddings for natural language processing[R]. Technical Report EPFL，2016.

[5] KIM Y，JERNITE Y，SONTAG D，et al. Character-aware neural language models[C]// Thirtieth AAAI Conference on Artificial Intelligence. 2016.

[6] KRUEGEL C，VIGNA G. Anomaly detection of web-based attacks[C]// 10th Conference on Computer and Communication Security. ACM，USA，2003：251-261.

[7] KUEGEL C，VIGNA G，ROBERTSON W. A multi-model approach to the detection of web-based attacks[J]. Computer Networks，2005，48(5).

[8] ROBERTSON W，VIGNA G，KRUEGEL C，et al. Using generalization and characterization techniques in the anomaly-based detection of web attacks[C]// Annual Network and Distributed System Security Symposium(NDSS). 2006.

[9] TEKEREK A，GEMCI C，BAY O F. Development of a hybrid web application firewall to prevent web based attacks[C]// 2014 IEEE 8th International Conference on Application of Information and Communication Technologies(AICT). 2014：1-4.

[10] APPLEBAUM S，GABER T，AHMED A. Signature-based and Machine-Learning-based Web Application Firewalls：A Short Survey[J]. Procedia Computer Science，2021，189：359-367.

[11] GAO Y，MA Y，LI D. Anomaly detection of malicious users' behaviors for web applications based on web logs[C]// 2017 IEEE 17th International Conference on Communication Technology (ICCT). 2017：1352-1355.

[12] SUNEETHA K R，KRISHNAMOORTHY K R. Identifying User Behavior by Analyzing Web Server Access Log File[J]. International Journal of Computer Science & Network Security，2009，9(4)：327-332.

[13] FENG Q Y. Research on Log Anomaly Detection and User Behavior Analysis based on Web Application [D]. Guangzhou：South China University of Technology，2019.

[14] LIANG J，ZHAO W，YE W. Anomaly-Based Web Attack Detection：A Deep Learning Approach[C]// Proceedings of the 2017 VI International Conference on Network Communication and Computing. 2017：80-85.

[15] JEMAL I，HADDAR M A，CHEIKHROUHOU O，et al. M-CNN：A New Hybrid Deep Learning Model for Web Security [C]// 2020 IEEE/ACS 17th International Conference on Computer Systems and Applications(AICCSA). Antalya，Turkey，2020：1-7.

[16] LE H，PHAM Q，SAHOO D，et al. URLNet：Learning a URL representation with deep learning for malicious URL detection [J]. arXiv：1802. 03162，2018.

[17] ITO M，IYATOMI H. Web application firewall using character-level convolutional neural network[C]// 2018 IEEE 14th International Colloquium on Signal Processing and Its Applications (CSPA). IEEE，2018：103-106.

[18] SEYYAR Y E，YAVUZ A G，ÜNVER H M. Detection of Web Attacks Using the BERT Model[C]// 2022 30th Signal Processing and Communications Applications Conference(SIU). Safranbolu，Turkey，2022：1-4.

[19] BOKOLO B G，CHEN L，LIU Q. Detection of Web-Attack using DistilBERT，RNN，and LSTM[C]// 2023 11th International Symposium on Digital Forensics and Security(ISDFS). 2023：1-6.

[20] TRAN A T，LUONG T D，PHAM X S，et al. Deep Models with Differential Privacy for Distributed Web Attack Detection[C]// 2022 14th International Conference on Knowledge and Systems Engineering(KSE). Nha Trang，Vietnam，2022：1-6.

[21] SAXE J，BERLIN K. eXpose：A Character-Level Convolutional Neural Network with Embeddings For Detecting Malicious URLs[J]. arXiv：1702. 08568，2017.

[22] WU J. Convolutional Neural Network with Character Embeddings for Malicious Web Request Detection[C]// 2019 IEEE Intl Conf on Parallel & Distributed Processing with Applications，Big Data & Cloud Computing，Sustainable Computing & Communications，Social Computing & Networking. IEEE，2019：622-627.

[23] WANG J，ZHOU Z，CHEN J. Evaluating CNN and LSTM for web attack detection[C]// Proceedings of the 2018 10th International Conference on Machine Learning and Computing. 2018：283-287.

[24] PAL R，CHOWDARY N. Statistical profiling of n-grams for payload based anomaly detection for HTTP web traffic[C]// Proceedings of the 2018 IEEE International Conference on Advanced Networksand Telecommunications Systems(ANTS). Indore，India，2018.

[25] KHREICH W，KHOSRAVIFAR B，HAMOU-LHADJ A，et al. An anomaly detection system based on variable N-gram features and one-class SVM[J]. Information and Software Technology，2017，91：186-197.

[26] MIKOLOV T，CHEN K，CORRADO G，et al. Effcient Estimation of Word Representations in Vector Space[J]. arXiv：1301. 3781，2013.

[27] MIKOLOV T，SUTSKEVER I，KAI C，et al. Distributed Representations of Words and Phrases and their Compositionality[J]. arXiv：1310. 4546，2013.

[28] HTTP DATASET CSIC 2010[OL]. http：//www. isi. csic. es/dataset/.

[29] AHMAD F Z. WAF Dataset [OL]. https：//github. com/faizann24/Fwaf-Machine-Learning-driven-Web-Application-Firewall.

WANG Li，born in 1987，Ph. D，engineer. Her main research interests include network technology and network security.