基于 CNN-LSTM-attention 的 XSS 攻击检测方法

郑松奕[1] 陈国良[1] 张裕祥[1] 蒋正亮[1]
ZHENG Songyi CHEN Guoliang ZHANG Yuxiang JIANG Zhengliang

摘 要 在基于深度学习 XSS 检测的研究中，传统的 CNN、LSTM、CNN-LSTM 模型在某些数据集上可能存在一些问题。例如，CNN 可能无法有效处理具有复杂空间结构的数据，而在处理具有较长时间序列的数据时，LSTM 可能会出现梯度消失或梯度爆炸的问题。为了解决这些问题，引入 attention 机制，结合 CNN 和 LSTM 模型（CNN-LSTM-attention）用于 XSS 攻击检测。CNN-LSTM-attention 结合了 CNN 和 LSTM 优势，并通过注意力机制来提高分类的准确性。实验表明 CNN-LSTM-attention 相比 CNN、LSTM、CNN-LSTM 算法在准确率上有较大的提升。

关键词 跨站脚本攻击；卷积神经网络；长短期记忆网络；注意力机制

doi：10.3969/j.issn.1672-9528.2024.05.010

0 引言

近十年来，互联网技术不断发展，网络攻击数量不断增长，种类也日益繁多。根据国家信息安全漏洞共享平台（CNVD）统计，Web 应用在安全性方面面临着越来越严峻的挑战。与此同时，深度学习方法被广泛应用在 Web 攻击的检测中并取得了不错的成效。

本文选择 OWASP TOP10 中跨站脚本攻击（XSS）作为检测对象进行研究，使用深度学习的方法建立神经网络模型 CNN-LSTM-attention，对 Web 流量日志数据进行特征识别提取来识别入侵行为。

XSS 是最常见的恶意脚本注入攻击之一，当攻击者将恶意 Javascript 载荷注入到由用户的浏览器执行的 Web 页面时，就会发生这种攻击。XSS 攻击可能导致劫持用户会话、错误信息、篡改页面、插入恶意内容、网络钓鱼攻击、控制用户的浏览器和访问业务数据，甚至可能导致主机被攻击，威胁到受害者的内网安全[1]。

对于 XSS 检测方法的研究从传统的机器学习到近年来火热的深度学习都取得了不错的成效。传统的机器学习检测方法包括朴素贝叶斯、支持向量机、决策树等[2-3]，其主要思想是先人工提取 XSS 攻击载荷特征将其构建为特征向量，然后使用机器学习算法对 XSS 攻击进行检测。但是，这些基于传统机器学习的方法特点是严重依赖特征提取工作，缺少客观性，难以检测经过编码混淆的 XSS 攻击载荷。针对传统机器学习检测方法的不足，深度学习检测方法成为新研究热点。文献 [4] 提出使用 CNN 进行 XSS 检测，通过使用卷积神经网络（CNN）方法对 XSS 脚本进行分类，并将其识别为恶意脚本，在特征创建过程中几乎只使用 XSS 脚本字符。实验结果表明，该方法提高了准确率并减少了误报率，论证了 CNN 在 XSS 检测领域的有效性。文献 [5] 使用 LSTM-PCA 模型检测 XSS，实验结果表明，LSTM 有助于找出关键信息，并将特定的相关数据存储在内存中，验证了 LSTM 在 XSS 检测的有效性。文献 [6] 使用了 LSTM-attention 进行 XSS 攻击检测，该方法利用 LSTM 模型提取上下文相关特征的能力进行深度学习，增加的注意力机制使模型提取更有效的特征，可以更有效地识别 XSS 攻击。文献 [7] 研究对比了 CNN、LSTM、CNN-LSTM 在漏洞检测中的应用，实验结果表明，深度学习的方法在漏洞检测领域有较好的表现，优于多层感知器（MLP）等传统方法。然而，传统的 CNN 和 LSTM 模型存在一些问题，例如，在处理具有较长时间序列的数据时，LSTM 可能会出现梯度消失或梯度爆炸的问题，而 CNN 则可能无法有效处理具有复杂空间结构的数据。

基于上述研究，为了解决这些问题，本文引入 attention 机制，结合 CNN 和 LSTM 模型用于 XSS 攻击检测。具体来说，首先使用 CNN 提取文本特征；然后将这些特征输入到 LSTM 中以便对文本进行序列建模；接着引入注意力机制，从所有隐藏状态中选择要集中关注的部分，提高分类的准确性。

1 CNN-LSTM-attention 检测方法

CNN-LSTM-attention 检测方法整体框架如图 1 所示。

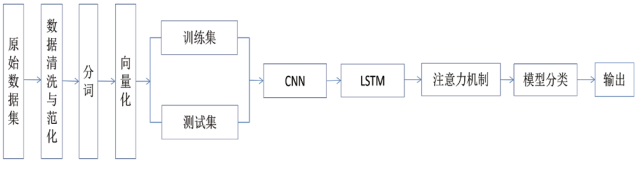

图 1 CNN-LSTM-attention 检测方法整体框架

1. 暨南大学网络与教育技术中心 广东广州 510632

首先收集相关的 XSS 数据作为原始数据集，然后对原始数据集进行预处理得到适应神经网络模型的输入，最后构建 CNN-LSTM-attention 混合模型进行检测分类 [8-10]。

1.1 原始数据集

本文所用的实验数据包含两个部分：恶意样例（正样例）来源于 XSSed 网站及笔者日常工作所收集的数据，有 41 734 条；正常样例（负样例）为 200 129 条正常的 Web 请求记录。本文实验中，全部数据随机切分为 70% 训练数据和 30% 测试数据。数据集的分布情况如表 1 所示。本文实验使用的计算机配置为：处理器 16vCPU，内存 32 GB，Centos Stream x64 操作系统。实验环境为 Python3.7、Keras 2.3.1、Tensorflow 2.2.0。

表 1 数据集分布

类别	训练集	测试集	总计
恶意样例	29 214	12 520	40 637
正常样例	140 090	60 039	200 129

1.2 数据预处理

数据预处理是将 XSS 样本转化为神经网络模型输入所需要的标准向量格式，即对收集到的数据进行解码、分词、向量化等数据预处理，然后输入到 CNN-LSTM-attention 模型（详见 1.3 节）中进行分类训练和测试。本文数据处理主要分为 3 个方面：数据清洗、分词、词向量化，完整的流程如图 2 所示。

图 2 数据预处理

1.2.1 数据清洗

攻击者往往会根据不同浏览器的解析策略对 XSS 攻击载荷进行编码混淆，以绕过检测，表 2 为常见的编码方式。为了提高模型的检测效果，本文根据表 2 的编码方式按照对应的编解码规则编写程序，对原始数据进行解码工作。实例结果见图 2 所示数据清洗部分的输出。

表 2 常见编码方式

编码名称	实例（原字符 / 编码）
URL	= / %3D
HTML	< / <
Unicode	= / \u003d

1.2.2 分词

对数据清洗后，为了降低向量化后的数据维度，首先将样本中的数字转换为 0，将链接如 http://xssed.org 或 https://xssed.org 转换为“http://u”。然后，根据 XSS 攻击载荷的特点，设计一系列正则表达式对样本数据进行分词，表 3 所示为部分分词正则表达式，分词结果见图 2 所示分词部分的输出。

表 3 分词

匹配类型	实例	正则表达式
开始标签	<script>、<h1> 等	<\w+>
结束标签	</script>、</h1> 等	</\w+>
http 链接	http://xssed.org 等	http://\w

1.2.3 词向量化

数据的向量化处理采用 Word2Vec 方法中的 Skip gram 模型，其模型是通过序列中一个词预测其周围词的向量。通过 Word2Vec 转换后的词向量不仅把词表示成分布式的词向量，而且可以捕获词之间存在的相似关系。神经网络的输入长度固定，但是样本长度不固定，选择合适的向量维度对模型的表现非常重要。根据样本长度，将长度超过向量维度的样本进行截断，长度不足的部分用 -1 填充，使得所有向量长度一致。实例结果见图 2 所示词向量化部分的输出。

1.3 CNN-LSTM-attention 模型

为了充分结合 CNN 和 LSTM 的特征提取特性，以及 attention 机制对关键特征的选取优势，本文采用 CNN 混合 LSTM 模型，并引入 attention 机制，提出 CNN-LSTM-attention 检测模型，如图 3 所示。

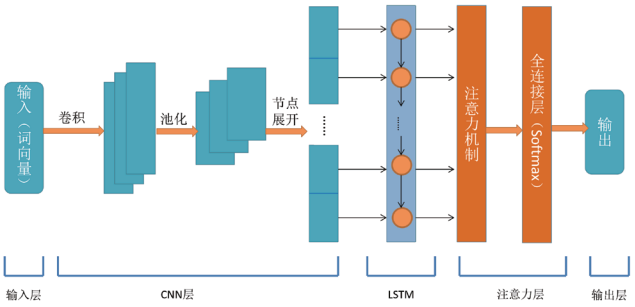

图 3 CNN-LSTM-attention 模型

输入层：数据预处理训练的词向量。

CNN 层：CNN 层可以提取数据中不同特征值之间的空间联系，进而弥补 LSTM 无法捕获数据空间分量的缺点，同时它提取出的特征仍然具有时序性。样本数据进入 CNN 层中会依次进行卷积、池化和节点展开（降维）操作。

LSTM 层：LSTM 具有记忆功能，可以提取出数据的时序变化信息，LSTM 隐藏层的输出会进入注意力层进一步处理。

注意力层：注意力可以提高 LSTM 中重要时间步的作用，从而进一步降低模型预测误差。注意力本质上就是求最后一层 LSTM 输出向量的加权平均和 [11-12]。LSTM 隐藏层输出向

量作为注意力层的输入，首先通过一个全连接层进行训练，然后对全连接层的输出使用 *Softmax* 函数进行归一化，最后得出每一个隐藏层向量的分配权重，权重大小表示每个时间步的隐状态对于预测结果的重要程度。权重训练过程为：

$$S_i = \tanh(WH_i + b_i) \quad (1)$$

$$\alpha_i = softmax(S_i) \quad (2)$$

利用训练出的权重对隐藏层输出向量求加权平均和，计算结果为：

$$C_i = \sum_{i=0}^{k} \alpha_i H_i \quad (3)$$

式中：H_i 为最后一层 LSTM 隐藏层的输出；S_i 为每个隐藏层输出的得分；α_i 为权重系数；C_i 为加权求和后的结果；*Softmax* 为激活函数。

输出层：输出预测结果。

2 实验结果及讨论

2.1 实验评估指标

实验用来评估 CNN-LSTM-attention 模型的指标包括准确率、精确率、召回率及 F_1 分数。

（1）准确率（accuracy）：指的是预测正确的结果占总样本的百分比，表达式为：

$$\text{准确率} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

虽然准确率能够判断总的正确率，但是在样本不均衡的情况下，并不能作为很好的指标来衡量结果。

（2）精确率（precision）：指的是在被所有预测为正的样本中实际为正样本的概率，表达式为：

$$\text{精确率} = \frac{TP}{TP + FP} \quad (4)$$

精确率和准确率看上去有些类似，但是二者为两个完全不同的概念。精确率代表对正样本结果中的预测准确程度，准确率则代表整体的预测准确程度，包括正样本和负样本。

（3）召回率（recall）：指的是在实际为正的样本中被预测为正样本的概率，表达式为：

$$\text{召回率} = \frac{TP}{TP + FN} \quad (5)$$

实际应用中以精确率还是以召回率作为评价指标，需要根据具体问题而定。

（4）F_1 分数：同时考虑精确率和召回率，让两者同时达到最高，取得平衡，是对召回率与精确率的一个综合评价。F_1 分数表达式为：

$$\text{F1分数} = 2 \times \frac{\text{精确率} \times \text{召回率}}{\text{精确率} + \text{召回率}} \quad (6)$$

以上公式中各变量的含义表示如下：TP（true positive）：将正样本预测为正样本数；FN（false negative）：将正样本预测为负样本数；FP（false positive）：将负样本预测为正样本数；TN（true negative）：将负样本预测为负样本数。

2.2 CNN-LSTM-attention 模型训练

2.2.1 向量维度

对于神经网络输入来说，选择合适的数据向量维度，才能充分利用样本信息。若向量维度过短，会遗失大量有效信息，降低检测准确率；若向量维度过长，则会增加训练的时间，不一定能提高准确率且降低检测实时性。为了得到合适的向量维度，本文比较了不同向量维度对准确率和训练时间的影响，结果如图 4 所示。实验结果表明，维度超过 48 时准确率有所下降，且训练时间不断增长，维度为 48 的准确率能达到最优，但是维度为 48 的训练时间明显低于维度超过 48 的训练时间，从而选择 48 作为输入数据的向量维度。

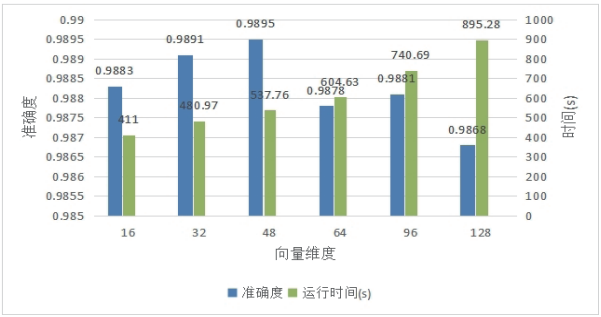

图 4 向量维度

2.2.2 CNN-LSTM-attention 模型参数设置

模型参数的设置会直接影响到模型的分类效果，本文中模型参数参考了文献[8-9]中参数设置，并根据实验进行调整，最终选择的 CNN-LSTM-attention 模型参数见表 4 。CNN 使用的是 keras 框架的 Conv1D 函数，卷积核通过参数 kernel_size 设置。

表 4 CNN 模型参数

参数	值
词向量维度	48
CNN 卷积核大小	kernel_size=3
LSTM 隐藏层大小	128
学习率	0.001
损失函数	交叉熵
优化函数	adam

图 5 展示了在当前参数设置下训练阶段损失（loss）和准确度（ACC）的变化。从图 5 可以看出，模型准确率稳步上升的同时，损失也在逐步下降，二者最终逐渐收敛到一个稳定值，这表明该模型具有良好的检测效果。

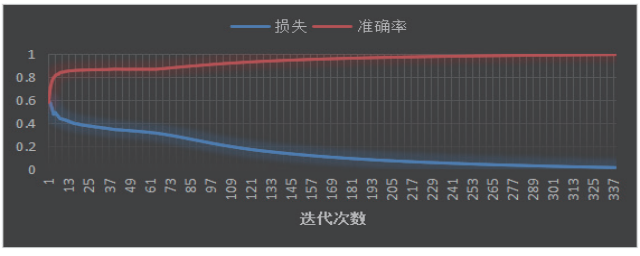

图 5 训练结果

2.3 CNN-LSTM-attention 模型测试

本文选取 CNN、LSTM、CCN-LSTM 模型进行对比，实验结果见表 5。可以看出，CNN-LSTM-attention 模型在准确率、召回率、F_1 分数方面均为最优，虽然准确率方面提升较小，但召回率和 F_1 分值有较大的提升。这是因为该方法首先通过使用 CNN 模型选取样本数据中最优的特征，然后经过 LSTM 捕获较长的依赖问题，最后通过 attention 选择相关性大的特征，进一步提高分类的准确率，所以 CNN-LSTM-attention 模型在准确率、召回率、F_1 分数方面有进一步提升。在时间收敛方面，CNN 收敛时间最短，LSTM 和 CNN-LSTM 收敛时间逐步增加，CNN-LSTM-attention 模型收敛时间最长，比 CNN-LSTM 增加了 1.82% 的收敛时间。

表 5 模型测试结果对比

模型	准确率 /%	召回率 /%	F_1 分数 /%	时间 /s
CNN	99.61	98.50	98.57	273.33
LSTM	99.59	98.69	98.80	323.11
CNN-LSTM	99.66	99.41	99.42	345.82
CNN-LSTM-attention	99.75	99.79	99.67	352.12

综上所述，CNN-LSTM-attention 模型在一定程度上能够有效地提取 XSS 中的全局和局部最优的特征，并且通过引入注意力机制，对不同的特征分配权重选取 XSS 中的有效特征，降低噪音特征对模型的干扰，从而进一步提高分类的准确率并减少训练时间。

3 结语

本文首先通过对数据集进行数据清洗、分词，再使用 word2vec 进行向量化形成输入样本数据。然后通过使用 CNN 提取 XSS 攻击载荷的特征，并使用 LSTM 建模序列数据。最后，使用注意力机制对不同数据分配不同权重，增加对关键信息的选取，减少对噪音特征的注意来提高分类的准确性。实验表明，本文提出的 CNN-LSTM-attention 模型相比 CNN、LSTM、CNN-LSTM 三种机器学习算法在准确率方面有进一步提升。相比未引入注意力机制的 CNN-LSTM 模型，CNN-LSTM-attention 模型在保证检测精度的同时仅增加了 1.82% 的收敛时间。

本文仅是使用 CNN-LSTM-attention 模型进行 XSS 漏洞检测，后期将研究该模型在 SQL 注入、跨站请求伪造等多种 Web 漏洞检测中的应用。

参考文献：

[1] NIRMAL K, JANET B, KUMAR R. It's more than stealing cookies-exploitability of XSS[C]//2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS). Piscataway:IEEE, 2018: 490-493.

[2] RATHORE S, SHARMA P K, PARK J H. XSSClassifier: an efficient XSS attack detection approach based on machine learning classifier on SNSs[J]. Journal of information processing systems, 2017, 13(4):1024-1028.

[3] WANG R, JIA X, LI Q, et al. Machine learning based cross-site scripting detection in online social network[C]//2014 IEEE Intl Conf on High Performance Computing and Communications, 2014 IEEE 6th Intl Symp on Cyberspace Safety and Security, 2014 IEEE 11th Intl Conf on Embedded Software and Syst. Piscataway:IEEE,2014: 823-826.

[4] NILAVARASAN G S, BALACHANDER T. XSS attack detection using convolution neural network[C]//2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF). Piscataway:IEEE, 2023: 1-6.

[5] STIAWAN D, BARDADI A, AFIFAH N, et al. An improved LSTM-PCA Ensemble classifier for SQL injection and XSS attack detection[J]. Computer systems science & engineering, 2023, 46(2):1759-1774.

[6] LEI L, CHEN M, HE C, et al. XSS detection technology based on LSTM-attention[C]//2020 5th International Conference on Control, Robotics and Cybernetics (CRC). Piscataway:IEEE, 2020: 175-180.

[7] WU F, WANG J, LIU J, et al. Vulnerability detection with deep learning[C]//2017 3rd IEEE international conference on computer and communications (ICCC). Piscataway:IEEE, 2017: 1298-1302.

[8] 李克资，徐洋，张庆玲，等. 基于 BiLSTM-attention-CNN 的 XSS 攻击检测方法 [J]. 贵州师范大学学报 (自然科学版), 2022, 40(4):76-83.

[9] 龚昕宇. 基于深度学习的 Web 攻击检测研究 [D]. 上海：上海交通大学，2020.

[10] 滕金保，孔韦韦，田乔鑫，等. 基于 LSTM-attention 与 CNN 混合模型的文本分类方法 [J]. 计算机工程与应用，2021, 57(14):126-133.

[11] 陈海涵，吴国栋，李景霞，等. 基于注意力机制的深度学习推荐研究进展 [J]. 计算机工程与科学 ,2021,43(2):370-380

[12] 刘建伟，刘俊文，罗雄麟. 深度学习中注意力机制研究进展 [J]. 工程科学学报 ,2021,43(11):1499-1511.

【作者简介】

郑松奕（1986—），男，广东揭阳人，硕士，助理工程师，研究方向：机器学习、软件工程、计算机网络与网络安全。

陈国良（1984—），通信作者（cgl@jnu.edu.cn），男，广东英德人，硕士，工程师，研究方向：网络管理、网络安全、系统管理、软件工程。

张裕祥（1989—），男，广东韶关人，硕士，工程师，研究方向：通信工程、网络安全。

蒋正亮（1987—），男，广东江门人，硕士，工程师，研究方向：计算机网络与网络安全。

（收稿日期：2024-02-21）