基于CNN与LSTM相结合的恶意域名检测模型

张 斌 廖仁杰*

(中国人民解放军战略支援部队信息工程大学 郑州 450001)

(河南省信息安全重点实验室 郑州 450001)

摘 要：为提高恶意域名检测准确率，该文提出一种基于卷积神经网络(CNN)与长短期记忆网络(LSTM)相结合的域名检测模型。该模型通过提取域名字符串中不同长度字符组合的序列特征进行恶意域名检测：首先，为避免N-Gram特征稀疏分布的问题，采用CNN提取域名字符串中字符组合特征并转化为维度固定的稠密向量；其次，为充分挖掘域名字符串上下文信息，采用LSTM提取字符组合前后关联的深层次序列特征，同时引入注意力机制为填充字符所处位置的输出特征分配较小权重，降低填充字符对特征提取的干扰，增强对长距离序列特征的提取能力；最后，将CNN提取局部特征与LSTM提取序列特征的优势相结合，获得不同长度字符组合的序列特征进行域名检测。实验表明：该模型较单一采用CNN或LSTM的模型具有更高的召回率和F1分数，尤其对matsnu和suppobox两类恶意域名的检测准确率较单一采用LSTM的模型提高了24.8%和3.77%。

关键词：恶意域名；卷积神经网络；长短期记忆网络；注意力机制

中图分类号：TN915.08; TP393 **文献标识码**：A **文章编号**：1009-5896(2021)10-2944-08

DOI: 10.11999/JEIT200679

Malicious Domain Name Detection Model Based on CNN and LSTM

ZHANG Bin LIAO Renjie

(*PLA Strategic Support Force Information Engineering University*, *Zhengzhou* 450001, *China*)

(*Henan Key Laboratory of Information Security*, *Zhengzhou* 450001, *China*)

Abstract: To improve the accuracy of malicious domain name detection, a new detection model based on Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) is proposed. The model extracts the sequence features from different length strings to classify the domain name. Firstly, in view of the sparseness of the N-Gram feature, the model utilizes CNN with different kernels to preserve the local association between the characters in the domain name strings and convert it to dense feature vectors. Secondly, in order to mine the context information of the domain name strings, LSTM is used to extract the deep-level sequence features of different character combinations. A sequence feature attention module is designed to assign little weight value to the sequence feature extracted from the padding characters, which decreases the interference by the padding characters and enhances the ability to capture distant sequence features. Finally, combining the advantages of CNN to extract local features and LSTM to extract sequence features, both partial and sequential information are put forward to improving the detection performance. Experimental results show that the recall rate and the F1-score of the proposed model are superior to other comparative models which are solely composed of CNN or LSTM. Particularly, when dealing with the matsnu and suppobox, the proposed model has increased by 24.8% and 3.77% in accuracy compared with the model based on LSTM, respectively.

Key words: Malicious domain name; Convolutional Neural Network (CNN); Long Short Term Memory (LSTM); Attention mechanism

收稿日期：2020-08-04；改回日期：2020-12-13；网络出版：2021-02-06

*通信作者： 廖仁杰 lrj2803@163.com

基金项目：河南省基础与前沿技术研究计划基金(142300413201)，信息保障技术重点实验室开放基金项目(KJ-15-109)，信息工程大学科研项目(2019f3303)

Foundation Items: The Foundation and Frontier Technology Research Project of Henan Province (142300413201), The Open Fund Project of Information Assurance Technology Key Laboratory (KJ-15-109), The Research Project of Information Engineering University (2019f3303)

1　引言

攻击者通过构造恶意域名达到诱导用户点击钓鱼网站、搭建命令与控制通道等目的，并且可结合Fast-Flux和Domain-Flux技术增强恶意域名躲避检测的能力[1]，如何准确、高效地检测恶意域名是网络安全领域亟需解决的问题。

基于合法域名和恶意域名在字符构成上的差异[2]，通常将恶意域名检测转化为对域名字符串的短文本分类研究。目前，基于域名字符串进行恶意域名检测的方法按特征提取的不同主要分为以下两类：第1类是基于手工特征提取的恶意域名检测方法，该方法从自然语言分析角度提取语言特征，如词素指标[2]，有意义字符占比[3]，N-Gram语法特征的KL散度、编辑距离与Jaccard系数[4]等作为域名特征，结合机器学习进行检测，此类方法对由随机字符拼接组成的恶意域名具有较好检测效果，但检测效果依赖特征工程，对不断更新的恶意域名变种需设计新的特征集，同时对通过单词拼接方式生成的恶意域名检测误报率较高；第2类是基于深度学习模型卷积神经网络(Convolution Neural Network, CNN)和长短期记忆网络(Long Short Term Memory, LSTM)的恶意域名检测方法：基于CNN的检测模型通过卷积核提取域名字符串中不同长度字符组的局部特征，设计CNN串联、并联[5,6]结构进行检测，此类模型检测速度快，但为提高恶意域名检测准确率，还需提取域名字符串深层次序列特征[7]；基于LSTM[8]的恶意域名检测模型通过提取域名字符串的序列特征进行域名检测，如以域名字符串的嵌入向量为输入的LSTM模型[9]和代价敏感的LSTM.MI[10]模型，比仅采用CNN的检测模型具有更高检测准确率，但仅考虑单字符序列特征，对单词拼接类恶意域名检测效果不佳。文献[11]设计CNN与LSTM相结合的恶意域名检测模型，采用混合词向量作为输入，通过结合域名字符串单字符和双字符序列特征提高模型多分类检测性能，但对单词拼接类恶意域名的检测准确率还需进一步提高。

综上，为提高深度学习模型检测恶意域名的准确率，考虑域名字符串中不同长度字符(单字符与多字符)组合的序列特征差异，在文献[11]的基础上，提出一种CNN与LSTM相结合的包含多条特征提取分支的恶意域名检测模型。其中，CNN用于提取域名不同长度字符组合的局部特征，LSTM用于提取不同长度字符组合局部特征的序列特征，同时引入注意力机制为LSTM不同位置输出的序列特征动态分配权值，通过特征加权降低填充字符对特征提取的干扰并增强序列特征提取能力。实验表明，结合域名不同长度字符组合的序列特征进行域名检测可有效提高恶意域名检测准确率，尤其是单词拼接类恶意域名的检测准确率。

2　基于CNN与LSTM相结合的恶意域名检测模型

2.1　模型组成

基于CNN与LSTM提取域名单字符和多字符组序列特征进行恶意域名检测的模型组成如图1所示。

该模型由输入层、特征提取层和输出层构成。其中，输入层将域名字符串转换为能被深度学习模型处理的数值向量，包含字符串长度补齐、one-hot编码和嵌入向量学习3个阶段，输出域名字符串的嵌入向量。

特征提取层包含LSTM层、1维卷积神经网络

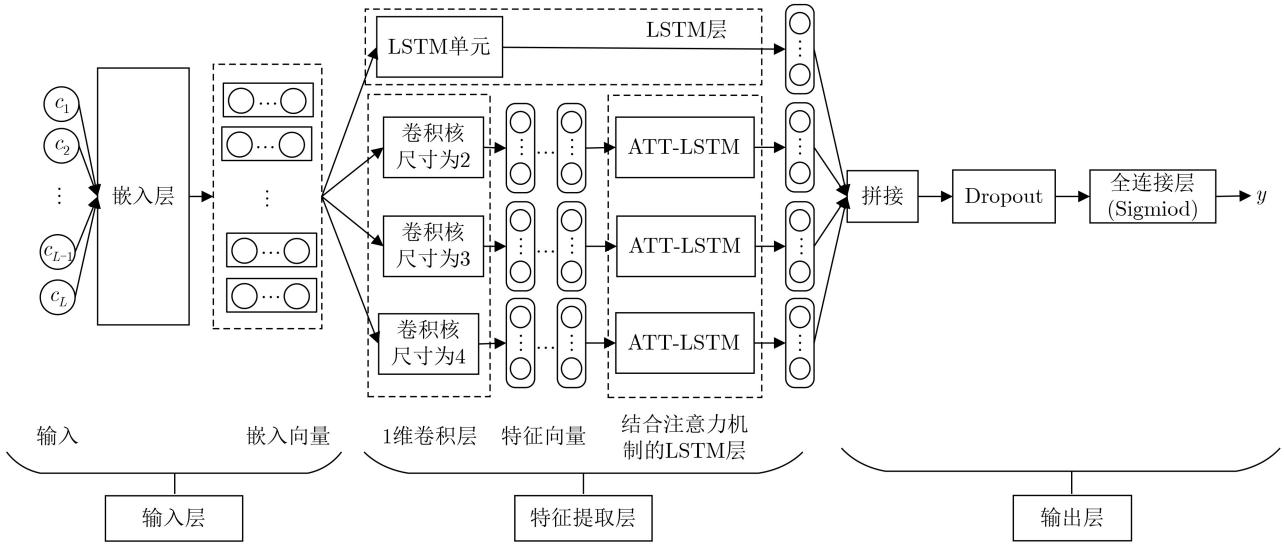


图 1 基于CNN与LSTM相结合的恶意域名检测模型(LSTM -Parallel CNN ATT-LSTM, L-PCAL)

(One-Dimension Convolution Neural Network，1D-CNN)层和结合注意力机制的长短期记忆网络(ATTention-based LSTM，ATT-LSTM)层，其中LSTM层用于提取域名字符串单字符序列特征；1维卷积神经网络层采用不同尺寸卷积核提取域名字符串中相邻字符局部特征，该层可看作域名字符串的N-Gram特征提取器；结合注意力机制的长短期记忆网络层采用LSTM提取1D-CNN输出的深层次序列特征并结合注意力机制降低填充字符对序列特征提取的干扰。

输出层将LSTM层输出特征与多个ATT-LSTM分支输出特征进行拼接，再经过Dropout层和激活函数为Sigmoid的全连接层处理，输出域名检测结果y。

2.2 输入层

本层的输入为2级域名标签。首先，将输入字符串长度统一为L，根据域名标签第1个字符必须为字母的限制，在长度不足L的域名标签前端填充“0”字符，得到长度为L的域名字符串，记为

$$S = c_1 c_2 \cdots c_L \tag{1}$$

其中，c_i为域名字符串中第i个字符，$c_i \in \{[a-z], [0-9], -\}$，$i \in \{1,2,\cdots,L\}$。然后，建立以域名字符串中字符为键，one-hot向量为索引值的字典D，并根据字典将域名字符串S转化为数值向量，记为

$$\mathbf{SD} = [\boldsymbol{D}(c_1)\ \boldsymbol{D}(c_2)\ \cdots\ \boldsymbol{D}(c_L)] \tag{2}$$

其中，$\boldsymbol{D}(c_i)$为字符c_i在D中的索引值，$i \in \{1,2,\cdots,L\}$。**SD**由one-hot向量组成，其特征维度大、特征值分布稀疏的特点将影响模型检测效果，采用Word2Vec技术[12]将**SD**转换为低维、稠密的嵌入向量，记为

$$\mathbf{SE} = [\boldsymbol{e}_1\ \boldsymbol{e}_2\ \cdots\ \boldsymbol{e}_L], \mathbf{SE} \in \mathbb{R}^{L\times d} \tag{3}$$

其中，$\boldsymbol{e}_i \in \mathbb{R}^d$为第$i$个字符的嵌入层向量，$i \in \{1,2,\cdots,L\}$，$d$为字符嵌入向量维度。

2.3 特征提取层

本层通过深度学习网络提取域名字符串的不同长度字符组合的序列特征，该层由LSTM层、1D-CNN层和ATT-LSTM层构成。

2.3.1 LSTM层

该层用于提取域名字符串单字符序列特征[9]，将域名字符串嵌入向量$\mathbf{SE} = [\boldsymbol{e}_1\ \boldsymbol{e}_2\ \cdots\ \boldsymbol{e}_L]$按$\boldsymbol{e}_1$至$\boldsymbol{e}_L$的顺序输入到LSTM单元进行特征提取，将输入$\boldsymbol{e}_L$后的输出作为域名字符串的单字符序列特征，记为

$$\boldsymbol{F}_{\mathrm{LSTM}} = \mathrm{LSTM}(\boldsymbol{h}_{L-1}, \boldsymbol{e}_L), \boldsymbol{F}_{\mathrm{LSTM}} \in \mathbb{R}^{d_{\mathrm{LSTM}}} \tag{4}$$

其中，LSTM()表示LSTM单元的特征运算，$\boldsymbol{h}_{L-1} = \mathrm{LSTM}(\boldsymbol{h}_{L-2}, \boldsymbol{e}_{L-1})$为输入$\boldsymbol{e}_{L-1}$时LSTM隐藏层输出，$d_{\mathrm{LSTM}}$为LSTM单元隐藏层输出向量维度。

2.3.2 1D-CNN层

域名字符串的N-Gram统计数据反映合法域名与恶意域名在不同长度字符组合的差异，是恶意域名检测的常用特征，但传统N-Gram特征存在随N值增大，特征维度指数增加、特征值稀疏分布的问题。为避免该问题，采用1D-CNN将域名字符串中相邻字符组合转化为维度固定、数值分布稠密的特征向量。考虑已有研究和计算开销，选取尺寸为2, 3, 4的卷积核提取域名字符串的2-Gram, 3-Gram和4-Gram特征。

对于域名字符串嵌入向量$\mathbf{SE} = [\boldsymbol{e}_1\ \boldsymbol{e}_2\ \cdots\ \boldsymbol{e}_L]$，1D-CNN通过尺寸为$k$，$k \in \{2,3,4\}$的卷积核$\boldsymbol{w}$，$\boldsymbol{w} \in \mathbb{R}^{k\times d}$进行特征提取包含以下两个步骤：卷积运算，将卷积核看作数据处理窗口，提取**SE**中与卷积核尺寸相同的数据与卷积核进行点积运算，再通过非线性激活函数运算得到对应位置的特征输出；卷积核移动，改变卷积核在**SE**的位置进行卷积运算得到不同位置的特征值，输出**SE**中长度为k的字符组合特征$\boldsymbol{x}$

$$\boldsymbol{x}\ :\ \boldsymbol{x}_j = f\left(\sum_{m=1}^{k}\sum_{n=1}^{d}(\mathbf{SE}[j+m,n]\cdot\boldsymbol{w}[m,n]) + b\right),\quad j \in \{1,2,\cdots,L-k+1\}, \boldsymbol{x} \in \mathbb{R}^{L-k+1} \tag{5}$$

其中，j为卷积核$\boldsymbol{w}$的位置参数，m与n为卷积运算参数，$\cdot$为乘运算，f为非线性激活函数，b为偏置项，$\boldsymbol{w}$与b通过模型训练进行更新。在实际中，采用多个尺寸相同但参数不同的卷积核进行特征提取，记宽度为k的卷积核个数为N_k，将N_k个卷积核所提取到的特征向量拼接，记为

$$\boldsymbol{F}_{\mathrm{CNN}}^{k} = [\boldsymbol{x}^1\ \boldsymbol{x}^2 \cdots \boldsymbol{x}^{N_k}], \boldsymbol{F}_{\mathrm{CNN}}^{k} \in \mathbb{R}^{N_k\times(L-k+1)} \tag{6}$$

其中，$\boldsymbol{x}^i$为第i个卷积核提取得到的长度为k的字符组合特征。

2.3.3 ATT-LSTM层

为利用域名字符串上下文语义信息进行域名检测，采用LSTM提取1D-CNN输出多字符组特征包含的序列特征。考虑输入层用于字符串长度填充的“0”字符对特征提取造成的干扰，引入前向反馈注意力机制[13]，通过引入注意力自学习函数动态为不同位置的LSTM单元输出分配权重，增强序列特征提取过程中抗干扰能力[14]，结合注意力机制的LSTM单元如图2所示。

图中特征向量为卷积核尺寸为k的卷积层提取得到的特征向量$\boldsymbol{F}_{\mathrm{CNN}}^{k} = [\boldsymbol{x}^1, \boldsymbol{x}^2, \cdots, \boldsymbol{x}^{N_k}]$，将$\boldsymbol{F}_{\mathrm{CNN}}^{k}$输入到LSTM单元得到不同位置的隐藏层输出，记为

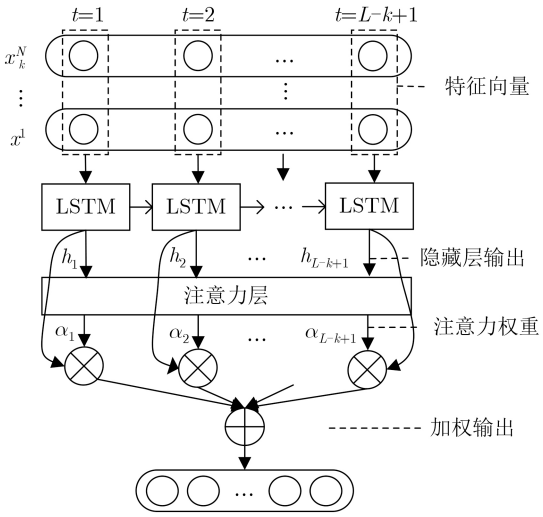


图 2 结合注意力机制的LSTM单元(ATT-LSTM)

$$\boldsymbol{h}:\boldsymbol{h}_t=\mathbf{LSTM}\left(\boldsymbol{h}_{t-1},\boldsymbol{F}_{\mathrm{CNN}}^{k}(:,t)\right),\quad t\in\{1,2,\cdots,L-k+1\}\tag{7}$$

其中，$\boldsymbol{h}_t$为输入$\boldsymbol{F}_{\mathrm{CNN}}^{k}(:,t)$时的隐藏层输出，$\boldsymbol{F}_{\mathrm{CNN}}^{k}(:,t)$为$\boldsymbol{F}_{\mathrm{CNN}}^{k}$中第$t$列数据构成的向量。为计算不同位置隐藏层输出的权重，引入自学习函数，记为

$$a\left(\boldsymbol{h}_t\right)=f\left(\boldsymbol{W}_\alpha\cdot\boldsymbol{h}_t+b_\alpha\right),t\in\{1,2,\cdots,L-k+1\}\tag{8}$$

其中，f为tanh函数，$\boldsymbol{W}_\alpha$为网络权重，b_α为偏置项，$\boldsymbol{W}_\alpha$和b_α通过模型训练进行参数更新。采用Softmax函数对权重进行归一化，得到归一化权重，记为

$$\boldsymbol{\alpha}:\alpha_t=\frac{\exp\left(a\left(\boldsymbol{h}_t\right)\right)}{\sum_{i=1}^{L-k+1}\exp\left(a\left(\boldsymbol{h}_t\right)\right)},t\in\{1,2,\cdots,L-k+1\}\tag{9}$$

将隐藏输出h与归一化权重$\boldsymbol{\alpha}$进行加权求和，得到字符数量为k的多字符组序列特征，记为

$$\boldsymbol{F}_{\mathrm{AttLSTM}}^{k}=\sum_{t=1}^{L-k+1}\boldsymbol{\alpha}_t\cdot\boldsymbol{h}_t\tag{10}$$

2.4 输出层

将3个ATT-LSTM分支输出的多字符组序列特征向量与LSTM层提取的单字符序列特征向量进行拼接，得到域名字符串的特征向量$\boldsymbol{F}$

$$\boldsymbol{F}=\left[\boldsymbol{F}_{\mathrm{LSTM}}\ \boldsymbol{F}_{\mathrm{AttLSTM}}^{2}\ \boldsymbol{F}_{\mathrm{AttLSTM}}^{3}\ \boldsymbol{F}_{\mathrm{AttLSTM}}^{4}\right]\tag{11}$$

将$\boldsymbol{F}$输入到激活函数为Sigmoid的全连接层，输出域名判别结果y，其中$y\in[0,1]$。在训练过程中加入Dropout层，即按照一定概率忽略域名字符串特征向量中部分位置的特征值，以降低模型训练过拟合的风险。采用交叉熵损失函数对检测模型训练效果进行量化，定义为

$$\mathrm{Loss}=-\frac{1}{N}\sum_{i=1}^{N}y_i^{'}\ln y_i+(1-y_i^{'})\ln(1-y_i)\tag{12}$$

其中，N为训练样本数量，$y_i^{'}$为样本真实标签，y_i为模型预测标签。

3 实验与结果分析

本节验证L-PCAL的有效性，选用Tensorflow 2.0和深度学习库Keras，GPU为NVIDA GTX 860M，使用Python语言实现检测模型搭建、训练与测试。

3.1 数据集

数据集中合法域名样本来自Alexa统计的域名列表[15]，随机选择35000条数据构成合法域名数据集；恶意域名样本来自360网络实验室[16]，考虑到不同恶意域名样本数量不均衡(如madmax类仅含1条数据，emotet类包含370747条数据)对模型训练和检测效果的影响，进行以下处理：对于样本数量大于2000的恶意域名类别，随机选取2000个样本作为实验样本，其次，删除样本集中数量小于5的恶意域名类别，最终构成包含41个类别，样本总数为35199的恶意域名数据集。将恶意域名与合法域名数据集合并后按5:1:1的数量比例拆分为训练集、验证集和测试集，其中合法域名标签设置为0，恶意域名标签设置为1。

3.2 模型对比与实验参数设置

选取Bi-Gram DT, LSTM[9], Bi-LSTM[5], Stack-CNN[5], Parallel-CNN[6], L-PCL, PCAL和CAL-PCAL作为L-PCAL的对比模型，其中Bi-Gram DT为基于域名字符串Bi-Gram特征的决策树检测模型、L-PCL为图1模型中去除注意力机制后构成的模型，PCAL为图1模型中去除LSTM层后构成的模型，CAL-PCAL为图1模型中将LSTM层替换为卷积核尺寸为1的卷积层与ATT-LSTM单元串联构成的模型。

实验参数设置如下：模型参数：LSTM隐藏层输出维度为64；CNN中非线性激活函数为线性整流函数(Rectified Linear Unit, ReLU)，均添加偏置项，不同尺寸卷积核数量均为32；Stack-CNN中卷积核尺寸为3和2；Parallel-CNN中卷积核尺寸为2, 3, 4；Dropout层丢失率为0.5，其余网络参数均为默认设置。输入层参数：域名字符串长度L为60，嵌入向量维度为32。模型训练：每轮迭代训练样本数量为100，为防止训练过拟合，将在验证集上取得最小损失值的模型作为最终模型。

3.3 评价标准

在测试集上进行模型性能评估，结果统计为以下4类：

TP(True Positive)：被正确判别为恶意域名的样本数；FP(False Positive)：被错误判别为恶意域名的样本数；TN(True Negative)：被正确判别为合法域名的样本数；FN(False Negative)：被错误判别为合法域名的样本数。实验参考的判别标准为

$$\text{查准率：Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}}$$

$$\text{查全率：Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}}$$

$$\text{误报率：FPR}=\frac{\text{FP}}{\text{FP}+\text{TN}}$$

$$\text{准确率：Accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}}$$

$$\text{F1分数：F1}-\text{Score}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$$

其中，查准率和误报率反映模型判别结果的可信度；查全率反映模型的漏报情况；F1分数反映模型综合性能，F1分数值越大说明模型性能越好。

3.4 测试集分类结果与分析

为验证L-PCAL的有效性，采用测试集测试不同模型的恶意域名二分类准确性，结果如表1所示，表中测试时间为模型检测10000条域名数据的处理耗时，各模型分别运行10次，最终结果取平均值。

由表1可知，Bi-Gram DT的各项检测性能指标与基于深度学习的检测模型差距较大，主要是由于Bi-Gram字符特征仅包含域名相邻距离为2的字符组特征且不包含域名字符串序列特征，从而导致较高误报率和较低的F1分数；由CNN构成的检测模型可提取域名字符串的局部特征，但需设计更深层的网络提取域名字符串长距离语义信息以提高检测准确性；由LSTM构成的检测模型可提取域名字符串序列特征，F1分数较由CNN构成的检测模型提升较大，其中，Bi-LSTM模型具有较低的误报率3.38%和较高的查准率96.49%，说明通过LSTM提取域名序列特征进行域名检测能有效提高检测准确率。

L-PCAL与L-PCL, PCAL, CAL-PCAL的对比实验采用控制变量的方法展开，L-PCAL与L-PCL的对比结果验证了引入注意力机制的有效性、与PCAL的对比结果说明了引入单字符序列特征的有效性、与CAL-PCAL的对比结果说明采用未结合注意力机制的LSTM用于单字符序列特征提取可取得较优的检测性能。L-PCAL结合域名字符串单字符序列特征和多字符组序列特征进行域名检测，具有最高的召回率93.91%和最高的F1分数0.9466。

在检测时间消耗方面，由于影响各模型运行耗时的因素包括模型参数规模、模型运算操作数、数据存取带宽等，故采用整体检测耗时作为模型检测效率的评价指标。Parallel CNN完成1次运算的操作数为62433次，测试耗时为0.57 s；LSTM完成1次运算的操作数为1491201次，测试耗时为4.46 s；L-PCAL为提高检测准确率，引入CNN与LSTM相结合的特征提取分支和注意力机制增加了模型的参数复杂度，其运算耗时主要集中在4个用于序列特征提取的LSTM单元，完成1次运算的操作数为5981408次，测试耗时为12.67 s。L-PCAL虽有较高计算复杂度，但其可在最高F1分数下达到789个/s的恶意域名检测速度，具有较高的检测效率和准确性。

3.5 误报率分析

检测模型的误报率和查全率的综合性能可采用观测者操作特性(Receiver Operating Characteristic, ROC)曲线进行衡量，其中ROC曲线与横坐标面积(Area Under Curve, AUC)越大，说明模型性能越好。在相同的测试集下，不同模型在恶意域名检测二分类任务中的ROC曲线如图3所示。

为刻画不同模型ROC曲线的差异，表2给出了不同误报率情况下，各模型的TPR和AUC对比结果。

由表2可知，L-PCAL在相同误报率情况下较对比模型具有最高的TPR和AUC值，说明模型通

表 1 模型检测性能对比表

模型	Recall (%)	Precision (%)	FPR (%)	F1-Score	Test Time(s)
Bi-Gram DT	84.37	75.32	22.60	0.7959	1.05
LSTM	93.75	93.58	6.57	0.9367	4.46
Bi-LSTM	90.88	96.49	3.38	0.9360	7.34
Stack-CNN	86.31	94.01	5.62	0.9001	0.62
Parallel-CNN	88.39	94.54	5.22	0.9136	0.57
PCAL	92.66	95.96	3.98	0.9428	12.16
L-PCL	92.17	96.38	3.54	0.9423	13.26
CAL-PCAL	93.02	95.41	3.98	0.9420	11.94
本文L-PCAL	93.91	95.42	4.61	0.9466	12.67

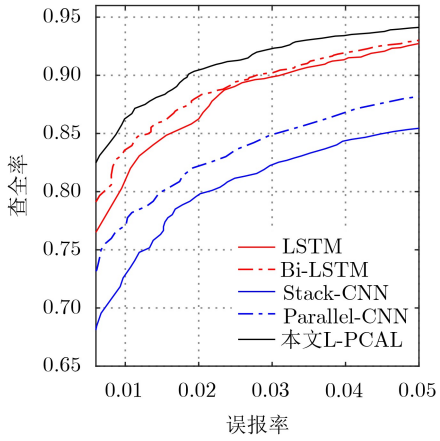


图 3 ROC曲线对比图

表 2　不同模型TPR与AUC对比表

模型	TPR (%)			AUC
	FPR: 1%	FPR:2%	FPR:3%	
LSTM	80.12	85.82	89.83	0.9846
Bi-LSTM	83.11	88.19	90.23	0.9840
Stack-CNN	72.58	79.82	82.24	0.9613
Parallel-CNN	77.13	82.04	84.85	0.9671
本文L-PCAL	85.74	90.40	92.17	0.9867

过多分支结构提取域名字符串不同长度字符组合的序列特征，可在较低误报率情况下有效区分合法域名与恶意域名，确保模型检测结果的可靠性。

3.6　单词拼接类恶意域名检测准确率对比与分析

攻击者为躲避不断升级的恶意域名检测技术，以mastnu和suppobox为代表的恶意程序从预存单词列表选取单词，通过单词拼接的方式生成恶意域名，如face-calculate.com。此类恶意域名在字符构成上与合法域名相似度较高，导致基于特征工程和基于深度学习的检测方法检测此类恶意域名误报率较高。

为验证L-PCAL检测单词拼接类恶意域名的检测性能，选取恶意域名样本中mastnu和suppobox构成单词拼接类恶意域名样本集，表3给出了不同深度学习检测模型在该样本集上的检测准确率。

对比表3和表1可知，基于深度学习的恶意域名检测模型对于单词拼接类恶意域名的检测准确率明显降低，尤其对于matsnu类别，部分深度学习模型检测准确率为0，说明此类模型无法区分单词拼接类恶意域名与合法域名的字符构成差异，从而将此类恶意域名误判为合法域名。单词拼接类恶意域名与合法域名在词语搭配方面存在差异，例如，合法域名“nationaljournal”为名词与名词搭配，而matsnu类的恶意域名“attempttrust”为动词与动词组合，语义信息难以理解。为有效检测单词拼接

表 3　单词拼接类恶意域名检测准确率对比表

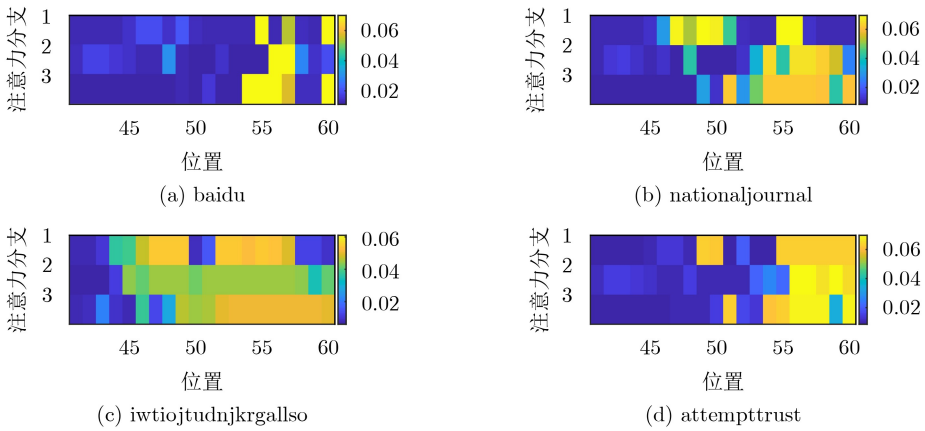
模型	Accuracy (%)	
	matsnu	suppobox
LSTM	0.78	81.57
Bi-LSTM	0.78	74.59
Stack-CNN	0	18.39
Parallel-CNN	0.78	16.08
PCAL	0	66.86
L-PCL	37.98	74.59
CAL-PCAL	7.75	74.59
本文L-PCAL	25.58	85.34

类恶意域名，L-PCAL首先通过不同卷积层提取域名不同长度字符组合的局部特征差异，再通过LSTM单元提取整体序列特征对域名做进一步区分，并引入注意力机制降低填充字符造成的干扰。L-PCAL在两类单词拼接类恶意域名样本检测中，检测准确率分别为25.58%和85.34%，较对比模型中表现最好的LSTM在两个类别的检测准确率分别提升24.8%和3.77%，有效提高单词拼接类恶意域名的检测准确率。

3.7　注意力可视化

ATT-LSTM层引入注意力机制的目的如下：(1)降低输入层用于长度填充的“0”字符对多字符组序列特征提取的干扰；(2)增大域名字符串中部分字符组合所处位置的序列特征在输出特征中所占权重。为验证引入注意力机制的有效性，选取合法域名baidu, nationaljournal和恶意域名iwtiojtud-njkrgallso，attempttrust作为L-PCAL输入，将ATT-LSTM单元中注意力结果输出并采用热力图进行可视化，如图4所示，图中每一行数据代表一条注意力分支在位置41～60的注意力取值。

由图4(a)可知，3个注意力分支权重在“baidu”字符串左侧填充字符“0”所处位置注意权值较小，故可有效降低在序列特征提取过程中填充字符造成的干扰。其次，合法域名字符串的权重仅在少数位置取得较大数值，如图4(b)中“nationaljourn-al”的权重在包含元音的字段“ourn”所处位置取得较大数值，而恶意域名字符串的权重较大值出现在随机字符组合对应位置和单词拼接类恶意域名的单词连接处，如图4(c)中字段“jkrgal”所处位置和图4(d)中“attempt”与“trust”连接处。通过注意力权值可视化一方面验证了引入注意力机制在消除干扰和增加部分字符组序列特征权重的有效性，另一方面为L-PCAL的序列特征提取过程提供一定的可解释性分析。



(a) baidu (b) nationaljournal

(c) iwtiojtudnjkrgallso (d) attempttrust

图 4 注意力权值可视化

4 结束语

本文提出一种基于CNN与LSTM相结合的恶意域名检测模型。该模型采用多个分支提取域名字符串中不同长度字符组合的序列特征刻画域名字符串的字符构成差异，并引入注意力机制降低特征提取过程中的噪声干扰。实验结果表明，所提模型能有效提高恶意域名检测准确率，尤其对恶意域名matsnu的检测准确率提高较大。下一步工作将引入模型压缩方法，以降低模型参数规模和模型运算时间。

参 考 文 献

[1] ZHAUNIAROVICH Y, KHALIL I, YU Ting, *et al*. A survey on malicious domains detection through DNS data analysis[J]. *ACM Computing Surveys*, 2018, 51(4): 67. doi: 10.1145/3191329.

[2] 张维维, 龚俭, 刘茜, 等. 基于词素特征的轻量级域名检测算法[J]. 软件学报, 2016, 27(9): 2348–2364. doi: 10.13328/j.cnki.jos.004913.

ZHANG Weiwei, GONG Jian, LIU Qian, *et al*. Lightweight domain name detection algorithm based on morpheme features[J]. *Journal of Software*, 2016, 27(9): 2348–2364. doi: 10.13328/j.cnki.jos.004913.

[3] SCHIAVONI S, MAGGI F, CAVALLARO L, *et al*. Phoenix: DGA-based botnet tracking and intelligence[C]. The 11th International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment, Egham, UK, 2014: 192–211. doi: 10.1007/978-3-319-08509-8_11.

[4] YADAV S, REDDY A K K, REDDY A L N, *et al*. Detecting algorithmically generated domain-flux attacks with DNS traffic analysis[J]. *IEEE/ACM Transactions on Networking*, 2012, 20(5): 1663–1677. doi: 10.1109/tnet.2012.2184552.

[5] YU Bin, PAN Jie, HU Jiaming, *et al*. Character level based detection of DGA domain names[C]. 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 2018: 1–8. doi: 10.1109/ijcnn.2018.8489147.

[6] SAXE J and BERLIN K. eXpose: A character-level convolutional neural network with embeddings for detecting malicious URLs, file paths and registry keys[EB/OL]. https://arxiv.org/abs/1702.08568, 2017.

[7] 杨路辉, 刘光杰, 翟江涛, 等. 一种改进的卷积神经网络恶意域名检测算法[J]. 西安电子科技大学学报, 2020, 47(1): 37–43. doi: 10.19665/j.issn1001-2400.2020.01.006.

YANG Luhui, LIU Guangjie, ZHAI Jiangtao, *et al*. Improved algorithm for detection of the malicious domain name based on the convolutional neural network[J]. *Journal of Xidian University*, 2020, 47(1): 37–43. doi: 10.19665/j.issn1001-2400.2020.01.006.

[8] HOCHREITER S and SCHMIDHUBER J. Long short-term memory[J]. *Neural Computation*, 1997, 9(8): 1735–1780. doi: 10.1162/neco.1997.9.8.1735.

[9] WOODBRIDGE J, ANDERSON H S, AHUJA A, *et al*. Predicting domain generation algorithms with long short-term memory networks[EB/OL]. https://arxiv.org/abs/1611.00791, 2016.

[10] TRAN D, MAC H, TONG V, *et al*. A LSTM based framework for handling multiclass imbalance in DGA botnet detection[J]. *Neurocomputing*, 2018, 275: 2401–2413. doi: 10.1016/j.neucom.2017.11.018.

[11] 杜鹏, 丁世飞. 基于混合词向量深度学习模型的DGA域名检测方法[J]. 计算机研究与发展, 2020, 57(2): 433–446. doi: 10.7544/issn1000-1239.2020.20190160.

DU Peng and DING Shifei. A DGA domain name detection method based on deep learning models with mixed word embedding[J]. *Journal of Computer Research and Development*, 2020, 57(2): 433–446. doi: 10.7544/issn1000-1239.2020.20190160.

[12] MIKOLOV T, CHEN Kai, CORRADO G, *et al*. Efficient

estimation of word representations in vector space[EB/OL]. https://arxiv.org/abs/1301.3781, 2013.

[13] RAFFEL C and ELLIS D P W. Feed-forward networks with attention can solve some long-term memory problems[EB/OL]. https://arxiv.org/abs/1512.08756, 2015.

[14] 谢金宝, 侯永进, 康守强, 等. 基于语义理解注意力神经网络的多元特征融合中文文本分类[J]. 电子与信息学报, 2018, 40(5): 1258–1265. doi: 10.11999/JEIT170815.
XIE Jinbao, HOU Yongjin, KANG Shouqiang, *et al*. Multi-feature fusion based on semantic understanding attention neural network for Chinese text categorization[J]. *Journal of Electronics & Information Technology*, 2018, 40(5): 1258–1265. doi: 10.11999/JEIT170815.

[15] Alexa Internet, Inc. Alexa top-ranked websites[EB/OL]. http://s3.amazonaws.com/alexa-static/top-1m.csv.zip, 2020.

[16] Qihoo 360 Technology Co, Ltd. 360 DGA feeds[EB/OL]. https://data.netlab.360.com/dga/, 2020.

张　斌：男，1969年生，教授，博士生导师，研究方向为信息系统安全.

廖仁杰：男，1996年生，硕士生，研究方向为基于机器学习的恶意域名检测.

责任编辑：马秀强