(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 109101552 B
(45)授权公告日 2022.01.28

(21)申请号 201810750707.2
(22)申请日 2018.07.10
(65)同一申请的已公布的文献号
申请公布号 CN 109101552 A
(43)申请公布日 2018.12.28
(73)专利权人 东南大学
地址 211189 江苏省南京市江宁区东南大学路2号
(72)发明人 杨鹏 曾朋 李幼平 张长江 郑斌
(74)专利代理机构 南京苏高专利商标事务所(普通合伙) 32204
代理人 李玉平
(51)Int.Cl.
G06F 16/955(2019.01)
G06N 3/04(2006.01)

(56)对比文件
CN 107992469 A,2018.05.04
CN 107169035 A,2017.09.15
CN 108009493 A,2018.05.08
US 2018063168 A1,2018.03.01

审查员 赵阳

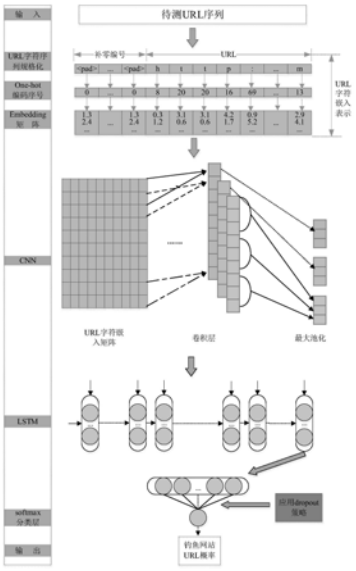
权利要求书3页 说明书5页 附图1页

(54)发明名称

一种基于深度学习的钓鱼网站URL检测方法

(57)摘要

本发明公开了一种基于深度学习的钓鱼网站URL检测方法，该方法仅根据网站URL就能够实时检测互联网上的钓鱼网站。本发明首先将URL字符串序列编码成one-hot二维稀疏矩阵，接着转化为稠密字符嵌入矩阵，输入到卷积神经网络中，抽取局部深度特征，然后将卷积神经网络的输出输入到长短期记忆网络，捕获URL序列的前后关联，最后接入softmax模型，对URL分类。本发明能避免繁冗的特征工程，通过卷积神经网络抽取局部深度关联性特征，通过长短期记忆网络学习URL中的长程依赖，能快速、准确地检测出钓鱼网站URL。



CN 109101552 B

1.一种基于深度学习的钓鱼网站URL检测方法,其特征在于,该方法主要包括三个步骤,具体如下:

步骤1,URL字符嵌入表示:首先将URL看做字符串序列,从字符层面量化URL,规格化URL,然后将URL字符转换成独热码(one-hot encode),最后通过卷积神经网络的嵌入(Embedding)层生成二维稠密矩阵即Embedding矩阵;

步骤2,CNN-LSTM分类层:Embedding矩阵首先通过卷积神经网络的CNN卷积层抽取局部关联性特征,接着抽取的局部关联性特征经池化层降低卷积神经网络模型复杂度;然后通过长短期记忆网络LSTM检测池化序列中的语义和长程依赖关系;最后输入到Softmax单元;

步骤3,模型训练:采用交叉熵损失函数,并利用Adam即自适应时刻估计算法迭代训练模型,优化损失函数;

步骤1中,URL字符嵌入表示将URL字符串序列量化编码,作为卷积神经网络CNN的输入;首先要确定URL中可能出现的所有字母字符、数字字符和特殊字符,并构建字符映射表;

假定每个URL字符序列长度固定为L,若URL长度超过L,则在URL末尾截取多余的字符,若URL长度少于L,则在URL首部补零直至长度达到L;

根据字符映射表,其中首部补零字符对应编号为0,URL中的字符“0”对应编号为53,最终每个字符被转换为长度为m的one-hot向量x,向量中字符对应编号位置为1,其余位置皆为0,因此URL被转换为公式(2)所示矩阵X;

$$X=(x_1,x_2,\ldots,x_L) \qquad (2)$$

将one-hot编码的矩阵X中的每个one-hot向量投影到d维连续向量空间$\mathbb{R}^d$;对应神经网络中的嵌入层,其可理解为一个输入为m个神经元,输出为d个神经元的全连接神经网络;

Embedding层的参数值随机初始化,并在模型训练过程中迭代更新;设输入为d个神经元,输出为m个神经元的Embedding全连接层参数矩阵为$W\in\mathbb{R}^{d\times m}$,则对one-hot向量$x_t$,$x_t$表示矩阵X的一个列向量,其最终嵌入向量$e_t$如公式(3)所示;

$$e_t=Wx_t=\begin{bmatrix} w_{11},w_{12},\ldots,w_{1m} \\ w_{21},w_{22},\ldots,w_{2m} \\ \vdots \\ w_{d1},w_{d2},\ldots,w_{dm} \end{bmatrix}\times\begin{bmatrix} x_{t1} \\ x_{t1} \\ \vdots \\ x_{tm} \end{bmatrix} \qquad (3)$$

最后URL字符串序列被转换为如公式(4)所示的稠密矩阵序列E,作为URL的字符嵌入矩阵;

$$E=WX=(w_1,w_2,\cdots,w_d)^T\times(x_1,x_2,\cdots,x_L)=(e_1,e_2,\cdots,e_L) \qquad (4)$$。

2.如权利要求1所述的基于深度学习的钓鱼网站URL检测方法,其特征在于,对步骤1中生成的URL字符嵌入矩阵E,将其输入到CNN-LSTM分类模型中,预测该URL为钓鱼网站的概率,步骤2实施过程分为3个子步骤:

子步骤2-1,卷积神经网络CNN层;CNN中卷积层对URL字符嵌入矩阵E进行卷积操作,抽取局部深度关联特征;具体而言,卷积层设置卷积核个数为S,每个卷积核都对窗口大小为k的字符嵌入向量进行卷积从而产生新特征;对于第f个卷积核,其在第i个滑动窗口处的字符向量矩阵E_i如公式(5)所示;

$$E_i=\{e_i,e_{i+1},\cdots,e_{i+k-1}\} \qquad (5)$$

则卷积核f在第i个滑动窗口处产生的新特征$\mathbf{h_i^f}$如公式(6)所示，其中σ是卷积层的非线性激活函数，采用relu激活函数，$\mathbf{W_f} \in \mathbb{R}^{k\times d}$和$b_f$分别为该卷积核权重和偏置项；

$$h_i^f = \sigma(W_f \cdot E_i + b_f) \tag{6}$$

设置卷积核滑动步长为1，则卷积核f遍历滑动窗口E_0到E_{L-k+1}后产生的特征图向量h^f如公式(7)所示；

$$h^f = \left\{h_1^f, h_2^f, \cdots, h_{L-k+1}^f\right\} \tag{7}$$

将S个卷积核产生的特征图堆叠，便可得到卷积层的序列矩阵H_S，如公式(8)所示，其中H_S的第i列$\mathbf{h_i} \in \mathbb{R}^{S\times 1}$；

$$H_S = \{h_1, h_2, \cdots, h_{L-k+1}\} \tag{8}$$

池化层对新的序列矩阵H_S进行最大池化操作，获取池化窗口p内的最大特征值，从而最大化字符特征表示；设置池化层步长与池化窗口相同，则对特征图向量h^f最大池化后的特征如公式(9)和(10)所示，其中$\mathbf{p_j^f}$为第j块最大池化的特征值，p^f表示池化后的向量，$N=\lceil (L-k+1)/p \rceil$；

$$p_j^f = Max(h_{(j-1)*p}^f, h_{(j-1)*p+1}^f, \cdots, h_{j*p-1}^f) \tag{9}$$

$$p^f = \left\{p_1^f, p_2^f, \cdots, p_N^f\right\} \tag{10}$$

最终，将S个池化向量堆叠，即可得到池化层的序列矩阵H_P，如公式(11)所示，其中H_P的第i列$p_i \in \mathbb{R}^{S\times 1}$；

$$H_P = \{p_1, p_2, \cdots, p_N\} \tag{11}$$

子步骤2-2，长短期记忆网络LSTM层；将池化序列矩阵H_P输入到LSTM神经网络中，其中p_i对应第i个时刻LSTM网络的输入，最终LSTM的输出隐藏状态序列H，如式(12)所示；

$$H = (h_1, h_2, \cdots, h_N) \tag{12}$$

接着将序列最后的隐藏状态h_N作为最后分类层的输入，如式(13)所示，其中n为LSTM网络隐藏单元个数，h_{Ni}为第i个隐藏单元；

$$h_N = (h_{N1}, h_{N2}, \cdots, h_{Nn}) \tag{13}$$

子步骤2-3，softmax分类层；分类层是激活函数为sigmoid的softmax回归单元，预测概率如式(14)所示，x为输入向量，w_k为权值向量，b_k为偏置，其中K＝2，当k＝0时，表示预测为正常网站的概率，k＝1时，表示预测为钓鱼网站的概率；

$$p(y=k|X) = \frac{\exp(w_k x + b_k)}{\sum_{i=0}^{K-1}\exp(w_i x + b_i)} \tag{14}$$

为了抑制过拟合现象，在隐藏状态h_N和softmax分类层之间的全连接层中应用dropout策略。

3.如权利要求2所述的基于深度学习的钓鱼网站URL检测方法，其特征在于，步骤3中模型训练的关键是确定目标损失函数，采用交叉熵损失函数，如式(15)所示；其中N为训练样本总数，y为样本的真实类别，0表示正常网站，1表示钓鱼网站，$\hat{y}$为模型预测为钓鱼网站的

概率；

$$L(\hat{y},y)=-\frac{1}{N}\sum_{i}^{N}[y_i\log\hat{y}_i+(1-y_i)\log(1-\hat{y}_i)] \tag{15}$$

采用自适应时刻估计算法训练模型优化交叉熵损失函数。

4.如权利要求1所述的基于深度学习的钓鱼网站URL检测方法，其特征在于，根据ASCCI码表并结合URL字符的实际情况，构建了97个编号的字符映射表，其中包括52个大小写字母a-Z，10个数字0-9，33个特征字符“—，；.!?:〞′/\|_@#$%ˆ&*～`+-＝<>()[]{}”，一个补零字符及未知字符编号。

一种基于深度学习的钓鱼网站URL检测方法

技术领域

[0001] 本发明涉及一种基于深度学习的钓鱼网站URL检测方法，该方法提取URL字符串序列相关特征，利用深度学习方法提高分类准确率，能实时检测互联网上的钓鱼网站，属于网络空间安全技术领域。

背景技术

[0002] 近年来，随着互联网的飞速发展，互联网体系结构在安全方面所存在不足日渐显露，网络钓鱼、网络犯罪、隐私泄露等各类安全问题越来越突出。没有网络安全就没有国家安全，网络空间安全已经成为世界各国必须共同面对和解决的难题。在各类网络安全问题中，网络钓鱼是一种通过社会工程学或其它复杂技术手段窃取网站用户个人信息的犯罪行为，目前网络钓鱼呈逐年上升趋势。

[0003] 当前主流钓鱼网站检测方法是基于机器学习的钓鱼网站检测方法，该方法将钓鱼网站检测视为一个二分类或聚类问题，首先根据钓鱼网站的URL结构及页面元素与正常网站的差异性提取特征，然后运用相应的机器学习算法达到钓鱼网站检测和防御的目的。常见的钓鱼特征有URL词汇特征、HTML特征、第三方网站特征等，根据所用特征的不同，又可分为基于URL特征的钓鱼网站检测和基于组合特征的钓鱼网站检测。其中基于URL特征的钓鱼网站检测方法不需要关注钓鱼页面，检测效率高，但不能全面反映URL的特点，准确率不高。

发明内容

[0004] 发明目的：针对当前日益增多的钓鱼网站和已有基于URL特征的钓鱼网站检测方法准确率不高、漏报率和误报率较高的问题，本发明提出一种基于深度学习的钓鱼网站URL检测方法，首先将输入URL字符串规格化为固定长度，然后通过字符映射表将其转化为One-hot编码序号，接着嵌入层(Embedding Layer)将其转为稠密矩阵作为URL字符序列的特征表示，之后输入到CNN网络抽取局部深度特征，并通过LSTM解决长程依赖问题，最后将LSTM最后一个时刻的输出输入到softmax单元，该方法能实时检测互联网的钓鱼网站，相比传统基于URL特征的钓鱼网站检测方法，不需要手动抽取特征，能全面反映URL特征点，而且能够显著提供钓鱼网站检测准确率。

[0005] 技术方案：一种基于深度学习的钓鱼网站URL检测方法，该方法涵盖钓鱼网站检测的全过程。该方法主要包括URL字符嵌入表示、CNN-LSTM分类模型和模型训练等过程，能够有效捕获URL字符序列中字符前后的关联和语义信息，有效解决传统基于URL特征的钓鱼网站检测方法不能全面反映钓鱼网站URL特征的问题，并且将卷积神经网络和长短期记忆网络模型应用于钓鱼网站检测，提高检测准确率和减少检测漏报率。该方法主要包括三个步骤，具体如下：

[0006] 步骤1，URL字符嵌入表示。首先将URL看做字符串序列，从字符层面量化URL，规格化URL，然后将URL字符转换成独热码(one-hot encode)，最后通过嵌入(Embedding)层生成二维稠密矩阵即Embedding矩阵。

[0007] 步骤2,CNN-LSTM分类层Embedding矩阵首先通过CNN卷积层抽取局部关联性特征,接着抽取的局部关联性特征经池化层降低卷积神经网络模型复杂度;然后通过长短期记忆网络LSTM检测池化序列中的语义和长程依赖关系;最后将LSTM最后一个单元的输出到Softmax单元。

[0008] 步骤3,模型训练。本发明采用交叉熵(Cross Entropy)损失函数,并利用Adam(Adaptive Moment Estimation)即自适应时刻估计算法迭代训练模型,优化损失函数。

[0009] 有益效果:

[0010] 1.URL字符嵌入表示不需要手动抽取特征,且不损失任何信息地表征了URL信息,能全面反映URL特点。

[0011] 2.CNN-LSTM分类模型能够有效捕获URL字符序列中字符前后的关联和语义信息,具有更高的准确率、更低的漏报率和误报率。

附图说明

[0012] 图1为本发明整体流程图,包括URL字符嵌入表示和CNN-LSTM分类。

具体实施方式

[0013] 下面结合具体实施例,进一步阐明本发明,应理解这些实施例仅用于说明本发明而不用于限制本发明的范围,在阅读了本发明之后,本领域技术人员对本发明的各种等价形式的修改均落于本申请所附权利要求所限定的范围。

[0014] 本方法具体实施步骤如下:

[0015] 步骤1,URL字符嵌入表示。URL字符嵌入表示将URL字符串序列量化编码,作为卷积神经网络CNN的输入。为此,首先要确定URL中可能出现的所有字母字符、数字字符和特殊字符,并构建字符映射规则。根据ASCCI码表并结合URL字符的实际情况,构建了97个编号的字符映射表,其中包括52个大小写字母,10个数字,33个特征字符,一个补零字符及未知字符编号。字符映射表如表1所示。

[0016] 表1字符映射表

[0017]

字符	编号
abcdefghijklmnopqrstuvwxyz	1-26
ABCDEFGHIJKLMNOPQRSTUVWXYZ	27-52
0123456789	53-62
—,;.!?:'"/|_@#$%^&*~`+-=<>()[]{}	63-95
补零字符	0
未知字符	96

[0018] 假定每个URL字符序列长度固定为L,若URL长度超过L,则在URL末尾截取多余的字符,若URL长度少于L,则在URL首部补零直至长度达到L,如公式(1)所示。其中URL_s为原始URL字符串,len(URL_s)表示其总长度,PAD为首部补零字符串,其长度len(PAD)=L-len

(URL_s)，URL_s[0:L-1]为URL_s前L个字符，URL_f为规格化后的输入字符串。

[0019] $$URL_f=\begin{cases}PAD+URL_s & ,len(URL_s)<L\\ URL_s & ,len(URL_s)=L\\ URL[0:L-1] & ,len(URL_s)>L\end{cases} \quad (1)$$

[0020] 根据字符映射表，其中首部补零字符对应编号为0，URL中的字符“0”对应编号为53，最终每个字符被转换为长度为m(97)的one-hot向量x，向量中字符对应编号位置为1，其余位置皆为0，例如字符“a”表示为(0,1,0,…0)。因此URL被转换为公式(2)所示矩阵X。

[0021] $X=(x_1,x_2,...,x_L)$ (2)

[0022] 由于one-hot编码的矩阵X含有很多0，会带来稀疏编码且维度过高的问题，且这种表示不同字符之间完全没有空间及语义关联性，信息量太少。可将其转换到字符嵌入的低维稠密特征空间中，本文将矩阵X中的每个one-hot向量投影到d维连续向量空间$\mathbb{R}^d$。对应神经网络中的嵌入层，其可理解为一个输入为m个神经元，输出为d个神经元的全连接神经网络。

[0023] Embedding层的参数值随机初始化，并在模型训练过程中迭代更新。设输入为d个神经元，输出为m个神经元的Embedding全连接层的参数矩阵为$W\in\mathbb{R}^{d\times m}$，则对one-hot向量$x_t$，$x_t$表示矩阵X的一个列向量，其最终嵌入向量$e_t$如公式(3)所示。

[0024] $$e_t=Wx_t=\begin{bmatrix}w_{11},w_{12},...,w_{1m}\\ w_{21},w_{22},...,w_{2m}\\ \vdots\\ w_{d1},w_{d2},...,w_{dm}\end{bmatrix}\times\begin{bmatrix}x_{t1}\\ x_{t1}\\ \vdots\\ x_{tm}\end{bmatrix} \quad (3)$$

[0025] 最后URL字符串序列被转换为如公式(4)所示的稠密矩阵序列E，作为URL的字符嵌入矩阵。

[0026] $E=WX=(w_1,w_2,\cdots,w_d)^T\times(x_1,x_2,\cdots,x_L)=(e_1,e_2,\cdots,e_L)$ (4)

[0027] 步骤2，CNN-LSTM分类模型。对步骤1中生成的URL字符嵌入矩阵E，将其输入到CNN-LSTM分类模型中，预测该URL为钓鱼网站的概率，该步骤实施过程分为3个子步骤：

[0028] 子步骤2-1，卷积神经网络CNN层。CNN中卷积层对URL字符嵌入矩阵E进行卷积操作，抽取局部深度关联特征。具体而言，卷积层设置多个卷积核S，每个卷积核都对窗口大小为k的字符嵌入向量进行卷积从而产生新特征。对于第f个卷积核，其在第i个滑动窗口处的字符向量矩阵E_i如公式(5)所示。

[0029] $E_i=\{e_i,e_{i+1},\cdots,e_{i+k-1}\}$ (5)

[0030] 则卷积核f在第i个滑动窗口处产生的新特征h_i^f如公式(6)所示，其中σ是卷积层的非线性激活函数，本文采用relu激活函数，$W_f\in\mathbb{R}^{k\times d}$和$b_f$分别为该卷积核权重和偏置项。

[0031] $$h_i^f=\sigma(W_f\cdot E_i+b_f) \quad (6)$$

[0032] 本发明设置卷积核滑动步长为1，则卷积核f遍历滑动窗口E_0到E_{L-k+1}后产生的特征图向量h^f如公式(7)所示。

[0033] $$h^f=\left\{h_1^f,h_2^f,\cdots,h_{L-k+1}^f\right\} \quad (7)$$

[0034] 将S个卷积核产生的特征图堆叠，便可得到卷积层的序列矩阵H_S，如公式(8)所示，其中H_S的第i列$\mathbf{h_i} \in \mathbb{R}^{S\times 1}$。

[0035] $H_S=\{h_1,h_2,\cdots,h_{L-k+1}\}$ (8)

[0036] 池化层对新的序列矩阵H_S进行最大池化(Max Pooling)操作，获取池化窗口p内的最大特征值，从而最大化字符特征表示。设置池化层步长与池化窗口相同，则对特征图向量h^f最大池化后的特征如公式(9)和(10)所示，其中p_j^f为第j块最大池化的特征值，p^f表示池化后的向量，$N=\lceil (L-k+1)/p \rceil$。

[0037] $$p_j^f = Max(h_{(j-1)*p}^f, h_{(j-1)*p+1}^f, \cdots, h_{j*p-1}^f) \quad (9)$$

[0038] $$p^f = \left\{p_1^f, p_2^f, \cdots, p_N^f\right\} \quad (10)$$

[0039] 最终，将S个池化向量堆叠，即可得到池化层的序列矩阵H_p，如公式(11)所示，其中H_p的第i列$p_i \in \mathbb{R}^{S\times 1}$。

[0040] $H_P=\{p_1,p_2,\cdots,p_N\}$ (11)

[0041] 子步骤2-2，长短期记忆网络LSTM层。将池化序列矩阵H_p输入到LSTM神经网络中，其中p_i对应第i个时刻LSTM网络的输入，最终LSTM的输出隐藏状态序列H，如式(12)所示。

[0042] $H=(h_1,h_2,\cdots,h_N)$ (12)

[0043] 接着将序列最后的隐藏状态h_N作为最后分类层的输入，如式(13)所示，其中n为LSTM网络隐藏单元个数，h_{Ni}为第i个隐藏单元。

[0044] $h_N=(h_{N1},h_{N2},\cdots,h_{Nn})$ (13)

[0045] 子步骤2-3，softmax分类层。分类层是激活函数为sigmoid的softmax回归单元，预测概率如式(14)所示，x为输入向量，w_k为权值向量，b_k为偏置，其中K＝2，当k＝0时，表示预测为正常网站的概率，k＝1时，表示预测为钓鱼网站的概率。

[0046] $$p(y=k|X)=\frac{\exp(w_k x+b_k)}{\sum_{i=0}^{K-1}\exp(w_i x+b_i)} \quad (14)$$

[0047] 为了抑制过拟合现象，在隐藏状态h_N和softmax分类层之间的全连接层中应用dropout策略。dropout是深度神经网络中一种防止过拟合的高效方法，其在训练过程中，对每个神经网络单元，按照一定的概率将其从网络中丢弃。

[0048] 步骤3，模型训练。模型训练的关键是确定目标损失函数，本发明采用交叉熵(Cross Entropy)损失函数，如式(15)所示。其中N为训练样本总数，y为样本的真实类别(0表示正常网站，1表示钓鱼网站)，$\hat{y}$为模型预测为钓鱼网站的概率。

[0049] $$L(\hat{y},y)=-\frac{1}{N}\sum_{i}^{N}[y_i\log\hat{y}_i+(1-y_i)\log(1-\hat{y}_i)] \quad (15)$$

[0050] 本发明采用Adam(Adaptive Moment Estimation)即自适应时刻估计算法训练模型优化交叉熵损失函数，其是对梯度下降算法的改进，通过计算梯度的一阶矩估计和二阶矩估计而为不同的参数设计独立的自适应性学习率，避免了学习率消失、收敛过慢或损失函数波动较大的问题，具有高效的学习效果。

[0051] 综上所述，本发明涉及的主要参数如表2所示。

[0052] 表2主要参数

[0053]

参数	含义	取值
L	URL 固定长度	200
m	字符映射表大小即 one-hot 编码长度	97
d	Embedding 层维度	64
S	卷积核数量	128
k	卷积层窗口大小	5
p	池化层窗口大小	3
n	LSTM 隐藏单元数量	128

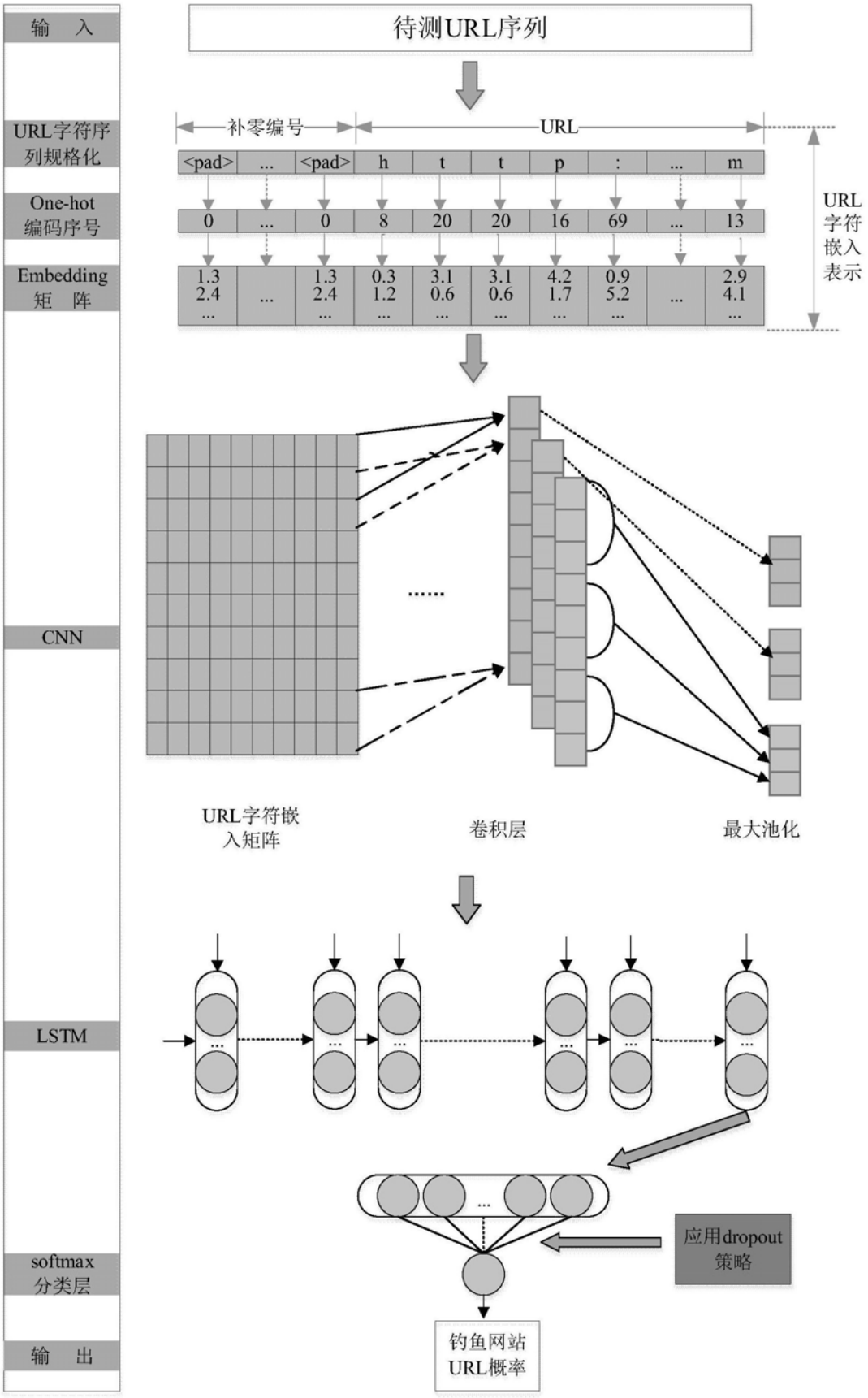

输 入
待测URL序列
URL字符序列规格化
补零编号
URL
<pad> ... <pad> h t t p : ... m
One-hot编码序号
0 ... 0 8 20 20 16 69 ... 13
Embedding矩 阵
URL字符嵌入表示
CNN
URL字符嵌入矩阵
卷积层
最大池化
LSTM
softmax分类层
应用dropout策略
输 出
钓鱼网站URL概率

图1