

基于双路特征提取和自适应融合的钓鱼网站检测

郭凯, 王恒, 马自强

(宁夏大学信息工程学院, 宁夏 银川 750021)

摘要: 随着网络钓鱼攻击手段的不断演进, 传统检测方法在特征提取、模型泛化能力等方面面临严峻挑战。针对这些问题, 提出了一种基于双路特征提取与自适应融合的钓鱼网站检测模型 (DPFEAT)。首先, 构建了时序与空间双路特征提取机制, 通过引入特征交互层增强整体特征的表征能力; 其次, 设计了自适应融合模块, 用于实现多维特征的动态融合, 从而有效提升模型对复杂攻击特征的捕获能力。在训练策略上, 采用预热-余弦衰减和循环学习率调度机制, 显著增强了训练过程的稳定性。同时, 结合 Focal Loss 与 Dice Loss 构建组合损失函数, 并引入动态权重平衡机制, 有效缓解了类别不平衡与边界样本识别困难的问题。此外, 还设计了可解释性分析框架, 用于评估各模块对模型性能的贡献。实验结果表明, DPFEAT 在准确率与受试者工作特征曲线下面积 (AUC-ROC) 值上分别达到 95.41% 和 98.83%, 显著优于现有方法。

关键词: 钓鱼网站检测; 双路特征提取; 自适应融合; 组合损失函数; 深度学习

中图分类号: TP393.08

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2025076

Phishing website detection based on two-way feature extraction and adaptive fusion

Guo Kai, Wang Heng, Ma Ziqiang

School of Information Engineering, Ningxia University, Yinchuan 750021, China

Abstract: With the continuous evolution of phishing attack methods, traditional detection approaches face severe challenges in terms of feature extraction and model generalization capability. To address these issues, a phishing detection model named DPFEAT (phishing detection model with dual-path feature extraction and adaptive fusion), which is based on dual-path feature extraction and adaptive fusion, was proposed. Firstly, a dual-path feature extraction mechanism for temporal and spatial features was constructed, and a feature interaction layer was introduced to enhance the overall feature representation capability. Secondly, an adaptive fusion module was designed to achieve dynamic fusion of multi-dimensional features, thereby effectively improving the model's ability to capture complex attack characteristics. Regarding the training strategy, a warmup-cosine decay and cyclical learning rate scheduling mechanism was adopted, which significantly improved the stability of the training process. Simultaneously, a combined loss function was constructed by integrating Focal Loss and Dice Loss, and a dynamic weight balancing mechanism was introduced, which effectively alleviated the issues of class imbalance and the difficulty in identifying boundary samples. Furthermore, an interpretability analysis framework was designed to evaluate the contribu-

收稿日期: 2025-03-05; 修回日期: 2025-12-21

通信作者: 王恒, wangh@nxu.edu.cn

基金项目: 国家自然科学基金 (No.62463025); 宁夏回族自治区重点研发计划一般项目 (No.2022BDE03008)

Foundation Items: The National Natural Science Foundation of China (No.62463025), General Project of Key Research and Development Program of Ningxia Hui Autonomous Region (No.2022BDE03008)

tion of each module to the model's performance. Experimental results demonstrated that the proposed DPFEAT model achieved an accuracy of 95.41% and an receiver operating characteristic area under the curve (AUC-ROC) score of 98.83%, which is significantly superior to existing methods.

Keywords: phishing website detection, two-way feature extraction, adaptive fusion, combined loss function, deep learning

0 引言

随着数字化进程的加速推进,网络钓鱼攻击已发展成为全球网络安全领域最严峻的威胁之一。网络钓鱼攻击指攻击者通过伪装成可信网站,非法获取受害者的用户名、密码、银行卡信息等个人敏感数据的犯罪行为^[1-2]。由于具备低成本、高回报与易实施等特征,此类攻击已成为网络犯罪分子最常用的攻击手段之一,对用户与企业均构成持续且严重的网络安全威胁。Proofpoint发布的《2024年网络钓鱼报告》显示,网络钓鱼攻击数量呈指数级增长,平均每月发生超过100万次网络钓鱼攻击。

传统的钓鱼网站检测方法主要依赖黑白名单、规则匹配等静态特征^[3],面对日益复杂和多样化的攻击手段,这类方法存在以下问题:①基于黑白名单的静态检测机制难以应对海量新出现的恶意域名;②规则匹配系统因攻击特征快速变异而失效;③基于单维度特征的机器学习模型在面对对抗样本时性能显著下降;④普遍存在误报率高的问题,严重影响企业安全事件处理效率;⑤可解释性的缺失使得安全分析师难以理解攻击模式、优化检测策略,也阻碍了有效的安全事件响应和取证溯源,削弱了用户和监管机构对自动化检测系统的信任。

深度学习技术的发展为上述问题提供了新的解决思路。然而,现有基于深度学习的钓鱼网站检测方法仍存在以下局限:①传统特征提取方式难以全面捕获钓鱼网站的多维特征;②模型结构简单,特征交互不充分,限制了模型的特征表达能力,从而影响识别精度;③训练过程不稳定,收敛效果不理想;④在真实的网络流量中,恶意钓鱼样本的数量远少于正常样本,传统模型在处理不平衡数据时表现不佳,往往倾向于准确预测多数类,而严重牺牲对关键少数类的识别能力,进而影响钓鱼检测效果。

针对现有方法的不足,本文提出了一种基于双路特征提取和自适应融合的钓鱼网站检测模型(DPFEAT)。

本文的主要贡献如下。

1) 设计了一种双路特征提取机制,分别从时序和空间维度深入挖掘网络钓鱼行为的特征,通过特征交互层增强特征表达能力,从而更准确地识别钓鱼攻击。

2) 设计了自适应融合模块,能够根据数据特性动态整合多维特征,显著提升模型对复杂钓鱼网站特征的捕获能力与模型的泛化性能。

3) 采用预热-余弦衰减的学习率调度策略,通过模拟余弦函数的衰减特性提高模型训练的稳定性和收敛效果,提升模型训练速度。

4) 结合Focal Loss与Dice Loss构建组合损失函数,配合动态权重平衡机制,有效缓解钓鱼网站检测中的少数类识别困难问题。

5) 设计了完整的可解释性分析框架,通过特征重要性分析、注意力权重可视化等解释机制,为模型的决策过程提供透明解释,增强模型的可信度。

1 相关工作

随着网络攻击手段的不断演进,研究者提出了多种基于不同技术的钓鱼网站检测方法。根据技术来划分,这些方法大致可以分为基于传统机器学习的钓鱼网站检测方法、基于深度学习的钓鱼网站检测方法、基于图像特征的钓鱼网站检测方法和基于多特征融合的钓鱼网站检测方法。

基于传统机器学习的钓鱼网站检测方法通常提取网站的特征用于训练分类器。顾晓清等^[4]提出了支持向量机(SVM)的钓鱼网站检测方法。该方法利用最优超平面对钓鱼网站和正常网站进行分类,具有较好的分类性能。尽管SVM能够有效处理高维特征数据,但其计算复杂度较高、对参数敏感、不直接支持概率输出、解释性较差,在某些场景中应用受限。Zhou等^[5]使用

LightGBM 等决策树模型结合多维度特征进行钓鱼网站检测,表现出较高的准确率和较低的误报率。虽然这类模型在检测精度上有一定优势,但其参数调优复杂,且可能存在过拟合风险,此外,基于树的模型通常缺乏对模型结果的解释,限制了其应用范围。

随着深度学习技术的发展,基于深度神经网络(DNN)的钓鱼网站检测方法逐渐成为研究热点。深度学习能够自动提取特征,并处理复杂的非线性数据,对提高检测准确率具有显著优势。Huang等^[6]提出基于卷积神经网络(CNN)的钓鱼网站检测方法,利用网页图像和文本特征进行分类。CNN能够自动提取高层次特征,特别适合处理复杂的网页数据,但对计算资源的要求较高,且模型的可解释性较差,难以对检测结果提供充分的解释。Wang等^[7]提出结合长短期记忆网络(LSTM)的钓鱼网站检测方法。该方法主要依赖对于序列数据(如网页内容的文本序列和用户行为序列)的处理。LSTM能够捕捉长期依赖性,但由于其对计算资源的要求较高且在小数据集上容易过拟合,应用范围受到限制。

近年来,图像处理技术也被引入钓鱼网站检测,尤其是在邮件钓鱼和网页仿冒检测方面。文伟平等^[8]提出一种利用EfficientNetV2模型进行网页视觉仿冒特性检测的方法,通过分析网页截图的Logo区域,识别图像中的仿冒特征,具有较高的识别率,但该方法需要较多的计算资源,且处理流程较复杂。戈晓洋等^[9]提出一种基于图像的钓鱼邮件检测方法,通过将邮件内容转化为图像并提取图像特征来识别钓鱼邮件。尽管该方法在准确率上有所提升,但在处理大规模数据时可能面临计算效率低的问题。张桥等^[10]提出基于消息传递注意力检测钓鱼统一资源定位符(URL)的方法。该方法通过敏感词分词对URL进行处理并转化为词嵌入向量,再利用基于消息传递注意力网络(MPAN)模型自动提取长距离、非连续的单词交互信息,实现对钓鱼网页的检测。

考虑到单一特征可能无法全面捕捉钓鱼网站的多样性,基于多特征融合的检测方法逐渐成为一种趋势。通过融合多个特征,如URL、超文本标记语言(HTML)内容、JavaScript代码等,能

够提高检测的准确率和鲁棒性。Ma等^[11]提出基于多特征融合和机器学习的URL恶意识别方法。该方法结合了URL、HTML、JavaScript等特征,并利用机器学习算法进行分类。尽管该方法在不同数据集上表现良好,但特征提取不足可能会影响最终的检测效果。张思睿等^[12]提出一种层级分类模型,通过融合域名信息、网页结构特征、网页文本内容及网页图标元素构建识别体系。该方法能够有效地应对攻击者的各种技术对抗策略,但对计算资源的要求较高,且模型的可解释性较差。

钓鱼网站检测技术已取得显著进展,从传统机器学习方法到深度学习方法,检测精度不断提高。本文研究聚焦于优化模型准确度、提升可解释性,并探索更加多元化的特征融合方式,以应对日益复杂的网络钓鱼威胁。

2 基于双路特征提取和自适应融合的钓鱼网站检测模型

2.1 模型总体架构

本文提出一种基于双路特征提取和自适应融合的钓鱼网站检测模型DPFEAT,其架构如图1所示。

该模型的核心设计思路如下。

1) 双路特征提取机制。时序特征分支专注于捕获URL和HTML文档中的序列模式,通过深度卷积网络和注意力机制提取时序依赖特征,有效识别钓鱼网站常见的时序模式;空间特征分支着重提取网页的结构化特征和视觉特征,采用空间注意力增强的卷积网络捕获页面布局、元素分布等空间关系。互补性设计使两个分支分别关注不同维度的特征,实现更全面的特征提取。

2) 特征交互增强设计。通过跨维度特征交互实现时序特征与空间特征的深度融合,利用注意力机制动态建立特征间关联关系,并在不同抽象层次上进行特征交互,提升模型的表达能力。

3) 自适应融合机制。根据特征的重要性动态调整融合权重;结合特征上下文信息进行自适应融合;在多个层次上进行特征融合,保留不同粒度的信息。

2.2 时序特征与空间特征

时序特征关注数据的时间依赖性 or 顺序信息

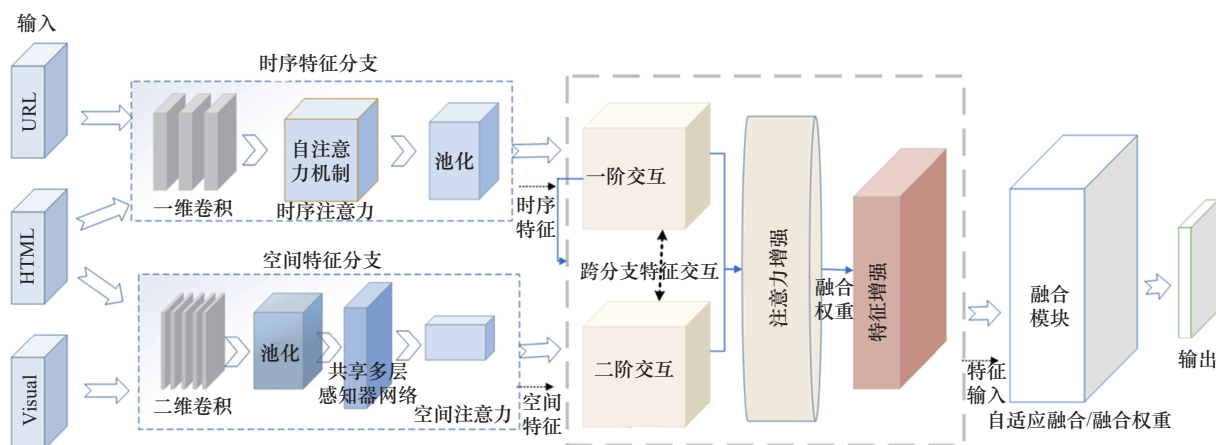


图1 DPFEAT模型架构

(如URL的字符序列、HTML标签的嵌套顺序和页面加载时间戳)，能够反映钓鱼网站对合法网站的模仿行为。

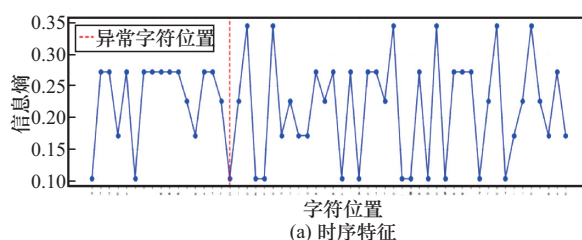
空间特征关注数据的结构或位置关系，能够揭示钓鱼网站在视觉设计上（如页面元素的布局坐标、视觉元素的空间分布和文档对象模型（DOM）节点的层级结构的仿冒行为。

HTML文档特征具有混合性，同时包含时序和空间信息。例如，<div>标签的嵌套顺序属于时序特征，而<input>表单的位置坐标属于空间特征。因此，在模型设计中，时序特征分支提取顺序信息，空间特征分支提取位置信息，共同实现对HTML特征的全面提取。特征分类见表1。

表1 特征分类

特征类型	元素示例
时序特征	URL字符序列、HTML标签流、页面加载时间戳
空间特征	视觉元素位置、DOM树结构、页面元素布局坐标

时序与空间特征可通过具体用例体现。例如，钓鱼网站可能通过模仿合法网站的字符顺序（时序特征）实施欺骗，而其页面布局则通过复制合法网站的视觉结构（空间特征）增强迷惑性。图2分别展示了基于钓鱼网站URL字符序列的时序特征和基于视觉元素位置的空间特征，以直观说明二者的区别。图2(a)以URL字符序列的熵值变化展示时序特征，可见位于红线位置的字符异常（原本的“o”被替换为“0”），域名拼写错误是钓鱼网站的典型特征；图2(b)以热力图的形式展示了空间特征。



(a) 时序特征



(b) 空间特征

图2 时序特征和空间特征

2.3 输入层

输入层接收原始特征输入，包括URL特征序列、HTML文档特征和页面视觉特征。DPFEAT模型的输入层设计如图3所示。

(1) URL特征序列处理

通过计算基础特征（如长度特征、数字数量、特殊字符数量）、判断是否使用超文本传输安全协议（HTTPS）、查询参数等，提取URL的多维度特征^[13]。长度特征为URL字符串的长度，数字数量为URL中数字字符（0~9）的出现次数，特殊字符数量为“-”“.”“/”“”“=”“&”等特殊字符的出现次数，是否使用HTTPS以“0”和“1”表示，查询参数指查询“?”后参数的字符数。将上述特征以数值形式编码，最终组成URL的基础特征向量，作为模型输入的一部分。

(2) HTML文档特征处理

从HTML内容中提取多个特征，有助于分析

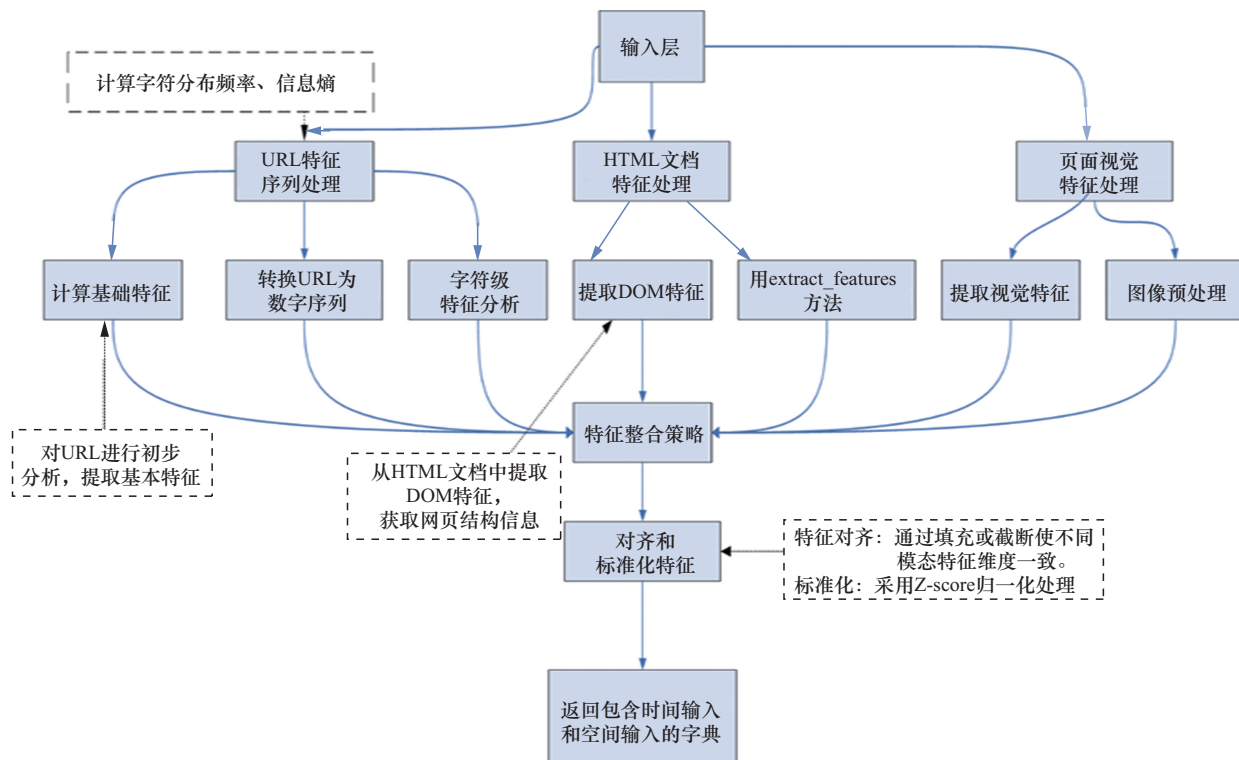


图3 DPFEAT模型的输入层设计

网页结构 and 安全性。本文采用特征提取（`extract_features`）方法提取DOM的特征，包括标签分布、DOM深度、总元素数、隐藏元素数及`<iframe>`、`<form>`、`<input>`和`<script>`标签的数量，还有外部和内部链接的数量。这些特征有助于识别网页的潜在风险，进一步提升对网络钓鱼网站的检测能力。

（3）页面视觉特征处理

从图像中提取多种视觉特征，主要包括颜色直方图、纹理特征、布局特征、Logo检测、品牌元素和视觉相似度特征^[14]。具体方法是接收图像并返回提取的特征，对图像进行预处理，确保符合模型的输入要求。此环节需要用到参数`target_size`，该参数用于指定目标图像大小，本文设置为(224, 224)。

（4）特征整合策略

特征整合策略用于整合URL、HTML和页面视觉3类特征。该策略的配置对象包含3个部分，即`url_config`、`html_config`、`visual_config`。在初始化时创建3个特征提取器实例，分别用于提取URL、HTML和页面视觉特征。整合特征时首先调用各特征提取器提取特征，随后对提取的特征

进行对齐和标准化处理，最后返回一个包含时间输入和空间输入的字典。

DPFEAT模型的输入层设计具有多特征融合、全面特征捕获和互补性信息提取等优势，使其能够从多个维度理解网页特征。

2.4 双路特征提取

有效的特征提取是提高钓鱼网站检测模型性能的关键。传统的特征提取方法通常无法全面捕获钓鱼网站的多维特征，导致检测效果不理想。钓鱼网站通常具有时序特征和空间特征（如页面布局、HTML元素等）。为充分捕捉这些特征，本文设计了时序特征分支和空间特征分支的双路特征提取机制，通过协同工作更全面地捕获钓鱼网站的多维特征信息，从而提升检测效果。利用这种机制，模型能够更全面地学习网站编码序列特征和页面布局等信息，显著提高钓鱼网站的检测准确率。

2.4.1 时序特征分支提取

时序特征分支专注于处理具有序列性质的特征信息。在钓鱼网站检测中，URL字符序列、HTML文档结构、JavaScript代码序列等都具有明显的时序特征，这些序列蕴含重要的编码序列

特征信息，可以用来区分异常的URL结构、可疑的HTML标签组合、恶意的脚本代码模式等。时序特征分支结构如图4所示。时序特征分支首先采用多尺度卷积进行初步特征提取，通过并行的不同大小（3,5,7）的卷积核，从多种序列数据中捕获多尺度的局部特征，其形式化表达式如下。

$$F_{t,k}^{(i)} = \sigma(W_{t,k}^{(i)} \cdot X + b_{t,k}^{(i)}), \quad k \in \{3,5,7\} \quad (1)$$

其中， $F_{t,k}^{(i)}$ 表示使用大小为 k 的卷积核得到的第 i 层特征， $W_{t,k}^{(i)}$ 表示相应的卷积核权重， X 表示输出序列， $b_{t,k}^{(i)}$ 为偏置项。为突出序列中的关键信息，研究人员设计了层次化的时序注意力机制，即局部注意力和全局注意力两个层次。

在时序特征分支中，局部注意力和全局注意力分别用于捕获不同尺度下的序列特征。局部注意力特征 $\alpha_{\text{local}} = \text{Softmax}\left(\frac{Q_1 K_1^T}{\sqrt{d_k}}\right) V_1$ ，全局注意力

$$\text{特征 } \alpha_{\text{global}} = \text{Softmax}\left(\frac{Q_g K_g^T}{\sqrt{d_k}}\right) V_g, \quad \text{其中 } Q、K、V$$

为注意力模块中的查询、键和值矩阵， d_k 表示键向量特征维度。为实现局部注意力特征和全局注意力特征的自适应融合，本文引入了自适应权重系数 β 。将局部注意力特征和全局注意力特征分别通过全连接层进行特征变换，然后将二者拼接后输入Sigmoid激活函数 σ ，得到权重系数 β ，具体见式(2)。

$$\beta = \sigma(W_\beta[\alpha_{\text{local}}; \alpha_{\text{global}}] + b_\beta) \quad (2)$$

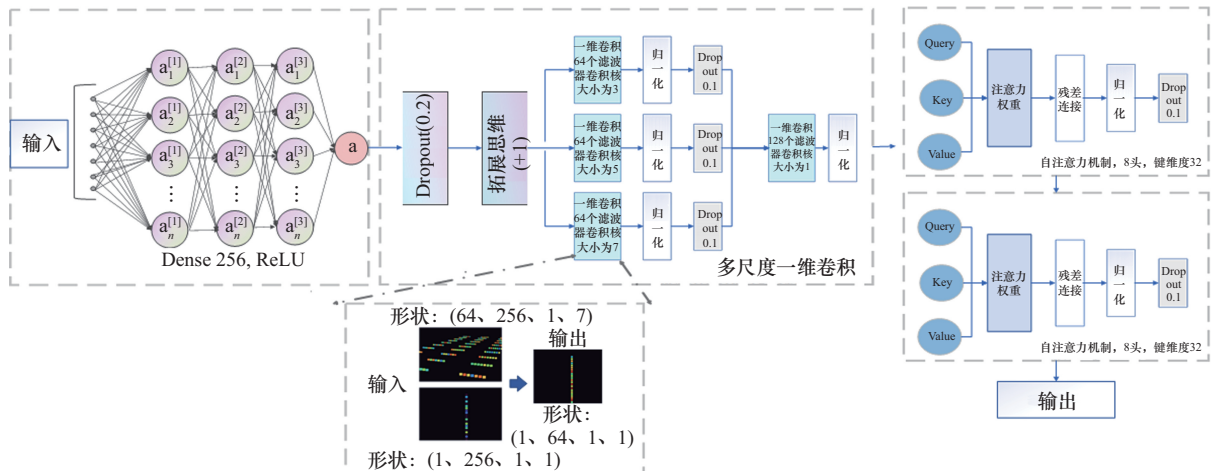


图4 时序特征分支结构

其中， W_β 和 b_β 均为可学习的参数。 β 在模型训练过程中通过反向传播自动学习和优化，实现对不同输入样本的自适应加权。最终的注意力特征 $F_{\text{att}} = \beta \alpha_{\text{local}} + (1 - \beta) \alpha_{\text{global}}$ 。

2.4.2 空间特征分支提取

空间特征分支主要处理网页的视觉和结构化特征（包括页面布局、元素分布、视觉样式等），这些特征对于识别仿冒正常网站的钓鱼页面尤为重要。空间特征分支结构如图5所示。本文模型采用深度可分离卷积进行空间特征提取，在降低计算复杂度的同时保持特征提取能力。

深度卷积层：

$$F_{s,\text{depth}}^{(1)} = \sigma(W_{s,d}^{(1)} \cdot F_s^{(1-1)}) \quad (3)$$

其中， $F_{s,\text{depth}}^{(1)}$ 是第1层的深度卷积输出， $W_{s,d}^{(1)}$ 是第1层的深度卷积核权重矩阵， $F_s^{(1-1)}$ 是第1-1层的输出。

点卷积层：

$$F_{s,\text{point}}^{(1)} = \sigma(W_{s,p}^{(1)} \cdot F_{s,\text{depth}}^{(1)}) \quad (4)$$

其中， $F_{s,\text{point}}^{(1)}$ 是第1层的点卷积输出， $W_{s,p}^{(1)}$ 是第1层的点卷积核权重矩阵。

在空间特征分支，研究人员设计了通道注意力和空间注意力的双重增强机制。通道注意力模块通过全局平均池化（global average pooling, GAP）和全局最大池化（global max pooling, GMP）提取特征图 F 的全局信息，将两个池化结果相加并输入多层感知机（multi-layer perceptron, MLP）中，生成通道注意力图 M_C ，Sigmoid函数将输出转换

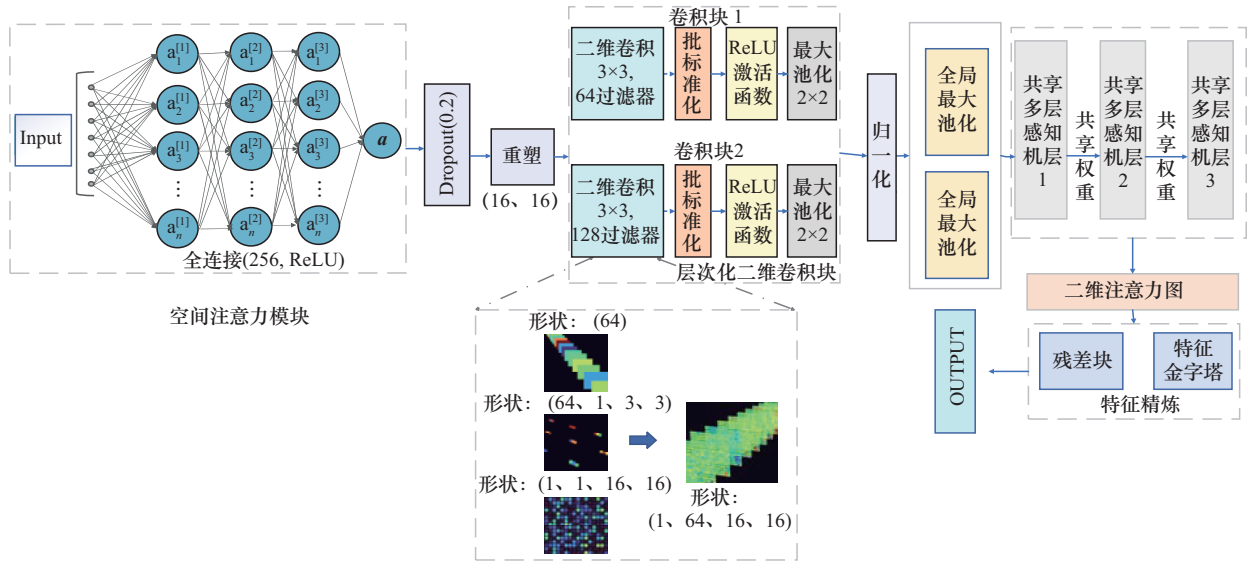


图5 空间特征分支的结构

为0到1之间的值。其形式化表示如式(5)所示。

$$M_C = \sigma(\text{MLP}(\text{GAP}(F) + \text{GMP}(F))) \quad (5)$$

其中, MLP 用于处理特征向量, $\text{GAP}(F)$ 表示将特征图 F 池化为一个全局平均特征向量, $\text{GMP}(F)$ 表示将特征图 F 池化为一个全局最大特征向量, $\text{GAP}(F) + \text{GMP}(F)$ 表示将两个池化结果逐元素相加。

空间注意力模块同样通过全局平均池化和全局最大池化提取特征图 F 的全局信息, 将两个池化结果在通道维度上拼接并输入 7×7 的卷积层中, 生成空间注意力图 M_S , 用 Sigmoid 函数将输出转化为0到1之间的值。其形式化表示如式(6)所示。

$$M_S = \sigma(\text{Conv}7 \times 7([\text{GAP}(F); \text{GMP}(F)])) \quad (6)$$

其中, $\text{Conv}7 \times 7$ 表示 7×7 的卷积操作, 用于处理特征图; $[\text{GAP}(F); \text{GMP}(F)]$ 表示将全局平均池化和全局最大池化的结果在通道维度上进行拼接。

将输入特征图 F 、通道注意力图 M_C 和空间注意力 M_S 逐元素相乘, 生成增强的特征图 F_{enhanced} 。其形式化表示如式(7)所示。

$$F_{\text{enhanced}} = F \otimes M_C \otimes M_S \quad (7)$$

通过这种方式, 模型可以更加关注重要的通道和空间位置, 从而提高特征表达能力和性能。

2.5 特征交互增强设计

2.5.1 交互层设计原理

特征交互增强层通过将时序特征和空间特征

拼接在一起, 形成一个更丰富的特征表示。该层首先将从时间序列中提取的特征 (形状为 $(\text{batch_size}, 128)$) 与从空间数据中提取的特征 (形状为 $(\text{batch_size}, 128)$) 在特征维度上合并, 形成一个新的特征张量 (形状为 $(\text{batch_size}, 256)$), 然后使用一个全连接层 (线性层) 对拼接后的特征进行处理, 将其映射回原始的特征维度 $(\text{batch_size}, 128)$ 。这种特征融合方法能够捕捉时序特征和空间特征之间的相互关系, 从而提升模型的表达能力和性能。

在特征交互方面, 本文设计了一个包含两个阶段的交互流程。

1) 一阶交互。首先进行直接特征交互, 即将时序特征和空间特征直接相乘, 生成一个新的特征表示。乘法操作可以捕捉时序特征和空间特征之间的交互信息。随后进行非线性变换增强, 即对交互后的特征应用全连接层进行非线性变换 (激活函数使用 ReLU), 引入非线性关系, 通过批量归一化稳定训练过程并加速收敛, 应用 Dropout (比例为 0.1) 防止过拟合, 增强模型的泛化能力。最后进行残差连接, 即将增强的特征与原始交互特征相加, 形成残差特征 (residual_features), 使模型更好地学习复杂特征, 同时减少信息丢失。

2) 二阶交互。首先对时序特征和空间特征分别使用全连接层进行投影, 将维度降为 64, 从而

获取更简洁的特征表示。随后进行外积运算，即将投影后的时序特征和空间特征融合，生成三维张量（形状为(batch size, 64, 64)）。这种方式能够捕捉时序特征和空间特征之间更复杂的交互关系。最后进行降维变换，即对外积结果进行重塑，将其转换为一维向量（形状为(batch size, 4096)），应用另一个全连接层（使用ReLU激活函数）将特征维度降为128，执行批量归一化以稳定训练过程。

交互特征融合的具体步骤如下：①将残差特征和二阶特征结合在一起，通过全连接层应用Softmax激活函数生成一个二分类的门控向量（gate）。此向量用于决定每个特征的权重。②通过乘法操作对两个特征进行加权（将残差特征乘以门控向量的第一维，将二阶特征乘以门控向量的第二维），从而可以根据门控的输出自适应调整特征的贡献。③通过加法操作将加权后的两个特征结合，生成最终的交互特征（interaction features）。这种融合方式充分利用了门控机制，可以有效提升模型对特征交互的学习能力，增强模型对输入信息的适应性和表达能力。

2.5.2 注意力增强机制

本文模型的注意力增强机制的具体实现过程如下。

1) 自注意力计算：首先通过一个全连接层（Dense）对输入特征进行处理，计算出注意力得分（attention scores），以此量化各输入特征的重要性；随后采用Softmax函数对注意力得分进行归一化处理，得到每个特征的权重分布，确保所有特征权重之和为1。

2) 注意力加权：将原始特征（features）和注意力得分（attention scores）相乘，生成加权

的特征（attended features）。该操作能够有效突出关键特征的贡献程度，同时抑制冗余或无关特征的干扰。

3) 多头注意力交互：引入Multi Head Attention实现特征的多头注意力交互。模型配置为8个并行注意力头，且每个注意力头的键维度设为16。通过多个注意力头的独立并行计算，模型可以从不同维度捕捉特征间的复杂关联，挖掘更丰富的语义与结构信息。

4) 残差连接与层归一化：将加权后的特征与多头注意力交互的输出相加，形成残差连接enhanced；通过层归一化（Layer Normalization）对残差连接的结果进行标准化处理，使训练过程更稳定。

2.5.3 交互特征增强

交互特征增强的实现过程如图6所示，具体步骤如下：首先，将输入的交互特征送入前馈神经网络（feed forward network, FFN），生成增强的特征表示。随后，对FFN的输出进行校准，确保特征分布更加平衡和有效。在此基础上，使用通道注意（channel attention）方法生成通道注意力，强调重要通道的特征。通道注意力通过加权不同通道的重要性来增强特征表达。类似地，采用空间注意（spatial attention）方法计算空间注意力，增强特征在空间维度上的表现，突出特征的空间分布和位置关系。最终，通过乘法操作将校准后的特征与通道注意力和空间注意力进行逐元素相乘，生成增强的特征。此设计通过联合通道与空间维度的注意力信息，显著提升了模型对特征的代表能力。

2.6 自适应融合设计

自适应融合模块能够根据数据特性动态整合多

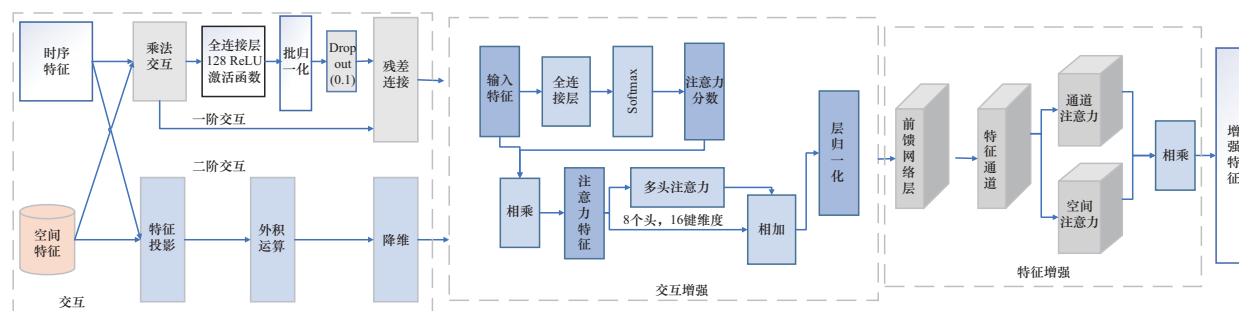


图6 交互特征增强的实现过程

维特征, 这些多维特征主要来源于以下两个方面。

1) 时序特征: 包括 URL 字符序列、HTML 文档结构中的标签序列、JavaScript 代码序列等具有时间顺序特性的信息。这类特征有助于捕捉钓鱼网站中可能存在的异常模式或恶意行为的时间依赖性。

2) 空间特征: 涵盖网页的视觉布局、元素分布等非时序但同样重要的结构性信息。例如, 页面上的图片、链接、表单, 以及其他可视组件的位置和排列方式, 可以帮助识别那些试图模仿合法网站外观的钓鱼页面。

这些多维特征通过特征交互增强后, 经自适应权重生成和加权特征融合进行整合。

2.6.1 自适应权重生成

自适应权重生成流程如下: 首先, 将各类输入特征进行拼接, 以整合其表征信息。随后, 通过 Dense(128, activation='relu') 计算交叉注意力, 输出的维度是 128, 采用 ReLU 激活函数引入非线性。在交叉注意力的输出上应用 Dropout(0.1),

随机丢弃 10% 的神经元, 防止过拟合。最终, 使用 Dense(3, activation='softmax') 计算融合权重, 输出维度为 3, 使用 Softmax 激活函数将输出转换为概率分布, 以此作为各输入特征的融合权重。

2.6.2 加权特征融合

本文设计了自适应融合模块实现加权特征的融合, 具体实现过程如图 7 所示。本文采用列表推导式对每个特征进行加权处理。首先通过 Multiply 层将每个特征与其对应的注意力权重相乘, 得到加权特征。随后, 利用 Add() 层将所有的加权特征相加, 融合为一个统一的特征表示 fused (即实现多源特征的聚合)。为进一步增强特征表达能力, 需对融合特征 fused 进行后续处理。首先通过 Dense(128, activation='relu') 层生成一个增强特征 enhanced, 其中的 ReLU 激活函数用于引入非线性。接着, 应用 Batch Normalization 对 enhanced 进行批量归一化, 加快训练速度并提高模型的稳定性。最后, 通过 Add() 层将原始的融合特征 fused 与增强特征 enhanced

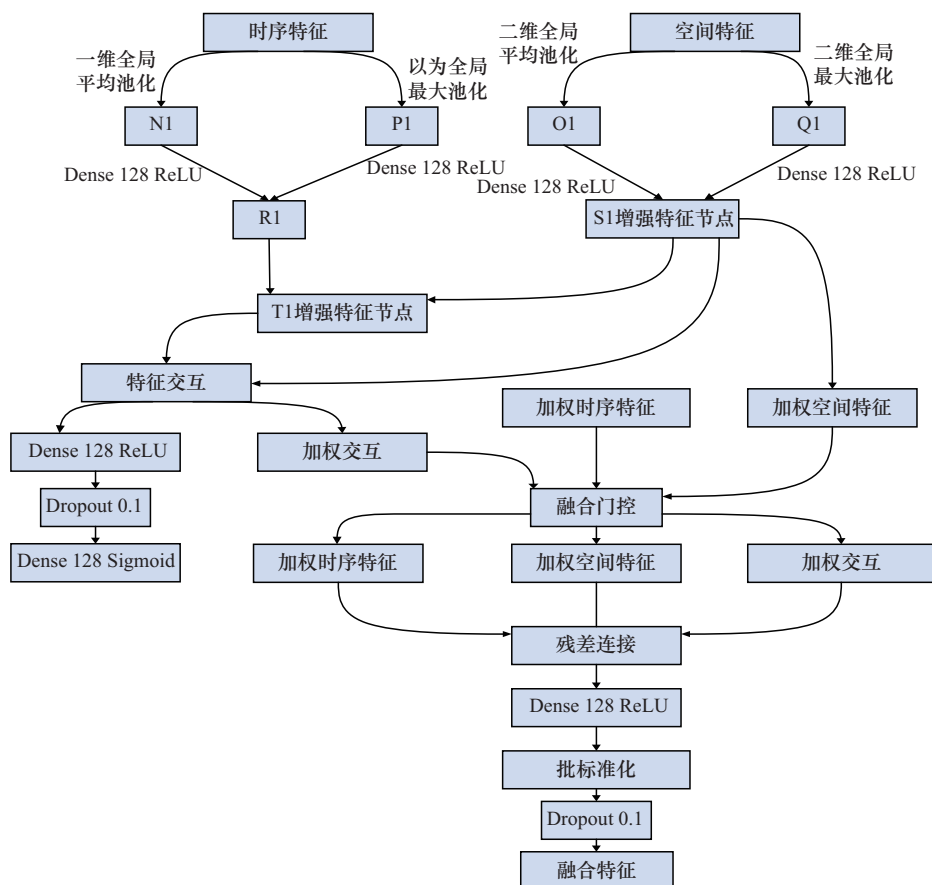


图7 自适应融合实现过程

相加，形成残差连接，从而保留输入信息和促进梯度流动。最终，函数返回一个融合且增强的特征表示 F_{fused} 。加权特征融合过程的形式化表示如式(8)所示。

$$F_{\text{fused}} = \sum_{i=1}^3 \alpha_i \odot F_i + \text{FFN}\left(\sum_{i=1}^3 \alpha_i \odot F_i\right) \quad (8)$$

其中， F_i 表示来自不同特征提取模块或不同层次的特征图； α_i 表示注意力权重，由注意力机制动态生成，用于调整每个特征图的重要性； $\odot$ 表示逐元素乘法，即将注意力权重与特征图进行逐元素乘法，生成加权特征图；加权求和用于将加权的特征图进行逐元素求和，生成综合特征图；FFN用于进一步处理综合特征图，增强特征的表达能力。

3 特征学习优化策略

为提升模型的性能与泛化能力，本文实施了多种特征学习优化策略，具体如图8所示。

3.1 梯度优化

梯度优化主要包括梯度裁剪与梯度累积。相关超参数（如梯度裁剪阈值）存储在配置（config）字典中。梯度优化具体实现流程如下。

1) 梯度裁剪：首先接收待处理的梯度列表，然后使用 TensorFlow 的 `tf.clip_by_global_norm` 函数对梯度进行裁剪，确保裁剪后梯度的全局范数不超过事先设定的最大值。这一操作能够有效避免训练过程中梯度值过大导致的“梯度爆炸”问题，保障模型训练的稳定性。

2) 梯度累积：其目的是通过小批次数据迭代模拟大批次训练效果。首先接收当前批次计算得到的梯度和预设的累积步数（steps）。在训练循环中，利用 TensorFlow 的 `tf.GradientTape` 记录每个批次的梯度计算过程，并计算当前批次数据对

应的梯度。这些梯度被依次累积到一个列表中。每完成一次梯度累加后，会将累积的梯度总和除以预设的累积步数，得到平均梯度。最终，模型使用该平均梯度进行参数更新。

3.2 正则化策略

在模型中定义 `RegularizationStrategy` 类，用于实现权重正则化和高级 Dropout 策略，防止过拟合。相关超参数（如 L2 正则化系数、Dropout 比率等）统一存储在配置（config）字典中。该策略的具体实现过程如下。

1) 权重正则化：首先遍历模型中所有的可训练权重，计算每个权重的平方和；随后通过 `tf.reduce_sum` 函数聚合所有可训练权重的平方和，得到总平方和；最后将总平方和与 config 中预设的 L2 正则化系数相乘，生成最终的正则化损失（`reg_loss`）并返回。

2) 高级 Dropout 策略：本文集成多种 Dropout 技术，构建了一套灵活适配的高级 Dropout 策略，增强模型的鲁棒性。具体实现逻辑如下。首先执行标准 Dropout 操作，根据模型训练状态（training 参数）激活相应的 Dropout 层。若输入为四维张量，则自动切换到 Spatial Dropout 2D 策略。该策略在特征图的空间维度上进行随机丢弃，既能有效去除冗余信息，又可保持特征图的空间关联结构不被破坏。若配置中启用了注意力 Dropout，则动态生成与输入维度匹配的随机注意力掩码，通过掩码与输入张量的逐元素相乘实现 Dropout。该策略在注意力模块中引入了可控随机性，有效提高了模型的泛化能力。

3.3 预热-余弦衰减和循环学习率的调度策略

在学习率调整方面，本文采用预热-余弦衰

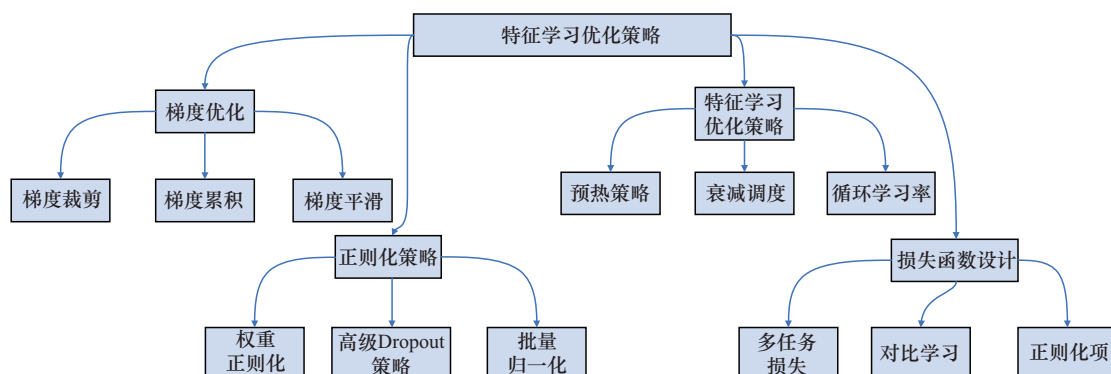


图8 特征学习优化框架

减学习率调度策略,并在此基础上引入循环学习率方法。在预热阶段,若当前训练步数小于预设的预热步数,则学习率按线性比例逐步提升,直至达到基础学习率。一旦预热阶段结束,进入余弦衰减阶段,学习率便依据余弦函数进行衰减。通过计算当前步数在总步数中的相对位置,并应用余弦函数公式求得学习率,确保学习率从最小值开始,平滑地衰减到最小值^[15]。循环学习率的实现基于当前训练周期,该周期由当前步数与步长共同决定。通过计算步长(记为 x),求得当前步数在周期内的位置,然后根据 x 的值计算学习率。学习率在最小学习率与最大学习率之间变化。当 x 达到最大值时,学习率回落至最小值。该机制有助于提升模型训练的稳定性与效率,从而促使模型能够更好地收敛。

在预热阶段,第 step 步时的学习率 $\text{lr}(\text{step})$ 可以通过式(9)求得。

$$\text{lr}(\text{step}) = \text{base_lr} \times \frac{\text{step}}{\text{warmup_step}} \quad (9)$$

其中, base_lr 表示预热阶段结束后使用的基础学习率, step 表示当前的训练步数(从0开始计数), warmup_step 表示预热阶段的总步数。当 $\text{step} < \text{warmup_step}$ 时,学习率从0逐渐增加到 base_lr 。当 $\text{step} \geq \text{warmup_step}$ 时,学习率保持为 base_lr 。

在余弦衰减阶段,第 step 步时的学习率 $\text{lr}(\text{step})$ 可以通过式(10)求得。

$$\text{lr}(\text{step}) = \min_lr + \frac{1}{2}(\text{base_lr} - \min_lr) \left(1 + \cos \left(\pi \times \frac{\text{step} - \text{warmup_step}}{\text{total_steps} - \text{warmup_step}} \right) \right) \quad (10)$$

其中, $\min_lr$ 表示训练结束时的学习率下限, total_steps 表示整个训练过程的总步数。当 $\text{step} < \text{warmup_step}$ 时,学习率从0逐渐增加到 base_lr 。当 $\text{step} \geq \text{warmup_step}$ 时,学习率按照上述余弦退火公式从 base_lr 逐渐下降到 $\min_lr$ 。

3.4 组合损失函数

本文采用结合 Focal Loss^[16]和 Dice Loss^[17]的组合损失函数,以应对样本不平衡问题。Focal Loss通过引入调制因子,增强模型对难分类样本的关注,从而减小易分类样本对损失的影响。而

Dice Loss则专注于优化预测结果与真实标签之间的重叠度,在处理类别不平衡的分割任务中表现优异。将二者结合,不仅能提升模型对难样本的学习能力、优化决策边界,还能增强模型在复杂数据场景下的鲁棒性与泛化能力。

计算 Focal Loss: $\text{focal} = \text{self.focal_loss}(y_true, y_pred)$ 。

计算 Dice Loss: $\text{dice} = \text{self.dice_loss}(y_true, y_pred)$ 。

计算组合损失: $\text{combined_loss} = (1 - \text{self.dice_weight}) \times \text{focal} + \text{self.dice_weight} \times \text{dice}$ 。

通过式(11)加权组合 Focal Loss 和 Dice Loss,通过式(12)求得最终的组合损失。

$$\begin{aligned} L_{\text{combined}} &= \lambda \text{Focal Loss}(p_i) + (1 - \lambda) \text{Dice Loss} \quad (11) \\ L_{\text{total}} &= \lambda \left(-\alpha_i (1 - p_i)^\gamma \log(p_i) \right) + (1 - \lambda) \left(1 - \frac{2 \sum p_i y_i + \epsilon}{\sum p_i^2 + \sum y_i^2 + \epsilon} \right) \quad (12) \end{aligned}$$

其中, ϵ 表述平滑系数,防止分母为零,增加数值稳定性; α_i 为类别平衡权重; y_i 为真实标签,表示第 i 个样本的真实类别; p_i 表示模型预测正确程度的概率; λ 是损失权重系数,用于控制 Focal Loss 和 Dice Loss 在组合损失中的相对重要性; γ 是 Focal Loss 中的聚焦参数,用于调整难易样本的权重。

动态权重调整:

$$\lambda(t) = \lambda_0 \cdot \exp(-\beta t) \quad (13)$$

其中, t 为时间变量,表示从初始时刻起经过的时间; $\lambda(t)$ 表示随时间 t 变化的权重值; λ_0 表示初始值,即 $t=0$ 时的值; β 为衰减率,用于控制衰减的速度,其值越大,表示衰减速度越快。

组合损失梯度:

$$\frac{\partial L_{\text{total}}}{\partial p} = \lambda \frac{\partial \text{FL}}{\partial p} + (1 - \lambda) \frac{\partial \text{DL}}{\partial p} \quad (14)$$

其中, L_{total} 代表总损失函数, p 代表模型预测概率。

4 实验及其分析

4.1 数据集

本文实验采用公开数据集 web-page-phishing-detection-dataset (版本号为v2024)。该数据集共包含 11 429 条样本,正类样本占比 48% (5 486 条),负类样本占比 52% (5 943 条)。该数据集涵盖

87个特征, 包括56个从URL结构与语法中提取的特征、24个来自页面内容的特征和7个通过查询外部服务获得的特征。训练集与测试集按5:1的比例划分。输入特征维度为256, 时序特征维度为128, 空间特征维度为16。在数据预处理阶段, 采用z-score标准化将特征值转换为均值为0、标准差为1的分布。缺失值使用均值填充, 从而确保数据的完整性。异常值检测阈值 $z\text{-score}=3.0$, 超过该阈值的样本被视为异常值。

4.2 评价指标

本文采用准确率 (Accuracy)、精确率 (Precision)、召回率 (Recall)、F1 分数 (F1-Score) 和 AUC-ROC (受试者工作特征曲线下面积) 5个指标评估模型性能。准确率反映模型的总体表现; 精确率与召回率聚焦于模型对正类样本的识别能力; F1 分数作为精确率和召回率的综合指标, 能够更好地平衡两者; AUC-ROC 用于衡量模型在不同分类阈值下区分正负样本的能力。

4.3 跨数据集泛化性验证

为验证 DPFEAT 的泛化能力, 研究人员在公开数据集 web-page-phishing-detection-dataset 和 3 个外部测试集 (Dataset Phishing Domain Detection || CyberSecurity、Phishing Dataset for Machine Learning 和 Phishing Websites Data) 上进行了实验, 结果见表 2。实验结果表明, DPFEAT 在 4 个数据集上均具有较高的分类精度 (平均 Accuracy 为 95.35%), 且 AUC-ROC 值稳定在 98% 以上, 说明模型在不同数据集上具有良好的泛化能力。

表 2 跨数据集泛化性实验结果

数据集	Accuracy	F1-Score	AUC-ROC
web-page-phishing-detection-dataset	95.40%	95.00%	98.30%
Dataset Phishing Domain Detection CyberSecurity	95.10%	95.20%	98.00%
Phishing Dataset for Machine Learning	96.00%	95.70%	98.40%
Phishing Websites Data	94.90%	95.10%	98.70%

4.4 对比试验

4.4.1 模型性能对比

研究人员将 DPFEAT 模型与机器学习模型 SVM、Random Forest、XGBoost、CatBoost、Gra-

dient Boosting、Decision Tree、LightGBM、RF+XGB、CatBoost + LightGBM、GradientBoosting+AdaBoost、DecisionTree+SVM 进行了对比实验, 性能对比结果见表 3。

表 3 DPFEAT 模型与机器学习模型性能对比结果

模型	Accuracy	Precision	Recall	F1-Score
SVM	85.30%	82.70%	89.20%	85.80%
Random Forest	89.00%	89.00%	88.70%	88.80%
XGBoost	89.20%	88.80%	89.80%	89.30%
CatBoost	89.50%	89.20%	89.70%	89.50%
Gradient Boosting	86.60%	85.90%	87.50%	86.70%
Decision Tree	85.00%	85.60%	84.50%	84.90%
LightGBM	89.00%	88.20%	90.10%	89.10%
RF + XGB	89.60%	89.10%	90.10%	89.70%
CatBoost+LightGBM	89.40%	88.90%	90.00%	89.40%
GradientBoosting+AdaBoost	86.60%	86.00%	87.50%	86.70%
DecisionTree+SVM	84.70%	84.90%	85.30%	84.70%
DPFEAT	95.40%	96.30%	95.50%	95.00%

从结果来看, 相较于其他模型, DPFEAT 的 Accuracy、Precision、Recall 和 F1-Score 均有提升。结果表明, 相比其他模型, DPFEAT 在自动特征提取方面性能更优, 这得益于 DPFEAT 的特征交互建模、时序依赖捕捉、空间关系建模能力, 以及有效的注意力机制。

DPFEAT 模型与主流深度学习模型性能对比结果见表 4。

表 4 DPFEAT 模型与主流深度学习模型性能对比结果

模型	Accuracy	F1-Score	AUC-ROC
DPFEAT	95.40%	95.00%	98.30%
CNN ^[18]	86.00%	86.80%	90.80%
Transformer (BERT) ^[19]	91.30%	91.70%	91.90%
DCGAN ^[9]	92.80%	无	无
ResNet ^[20]	92.50%	94.30%	94.20%
RNN ^[21]	84.00%	86.00%	85.00%
CNN-LSTM ^[22]	87.00%	86.00%	88.00%
DenseNet ^[23]	92.50%	94.30%	89.70%
CATALOG ^[24]	92.00%	91.00%	无

DPFEAT 模型的 Accuracy、F1-Score 和 AUC-ROC 值分别为 95.4%、95.0% 和 98.30%, 均优于其他模型, 表明 DPFEAT 具有较强的泛化能力。

对实验结果进行深入分析后发现, RNN 在处理长序列依赖问题上具有优势, 而 CNN 在处理短序列文本或捕捉局部特征方面更加高效。CNN 能够高效提取局部字符级特征, 而 RNN 需依赖循环结构处理时序依赖, 二者在 URL 特征处理上各有侧重。在本文中, CNN 通过多尺度卷积有效提取局部时序特征, 且具备更强的并行计算能力, 因此在 Accuracy 等指标上优于 RNN。

DenseNet 在 Accuracy 与 F1-Score 上与 DPFEAT 接近, 说明 DenseNet 对主流类别网页特征的识别能力较强。但 DenseNet 的 AUC-ROC 值较小, 表明其对少数类别网页特征的识别能力有限。DPFEAT 模型借助多分支特征融合与自适应加权机制, 提升了对少数类别的检测能力, 因此在 AUC-ROC 指标上表现更优。DCGAN 主要用于图像生成任务, 其网络结构与损失函数难以直接适配 URL/HTML 等序列化文本特征, 导致部分实验无法获得有效结果, 相关指标数据缺失。传统机器学习方法在特定数据集与特征工程下可能优于深度学习方法 CNN, 因 CNN 在小规模或特定领域数据上若缺乏足够的数据支持, 其性能无法充分发挥。CATALOG 先利用双重交叉注意力机制和双向掩码语言 (MLM) 模型融合局部与全局特征, 再输入多头自注意力模型; 同时还引入了时序特征。DPFEAT 相较于 CATALOG, Accuracy 提升超过 3%, 说明其时序-空间特征提取机制对网页的复杂特征具有更强的捕获能力。

此外, 研究人员还将 DPFEAT 模型与 Phishpedia^[25] 模型进行了对比。根据文献[25], 在品牌仿冒检测场景下, Phishpedia 的 Precision 达到 98.2%、Recall 为 87.1%。其 Recall 低于 DPFEAT (95.5%)。在经 PhishCatcher 筛选后的 URL 数据

集上, Phishpedia 的 Precision 降至 87.5%、Recall 为 87.5%。这表明 PhishCatcher 所筛选的 URL 中包含大量无品牌特征的钓鱼页面, 超出了 Phishpedia 基于 Logo 匹配的检测能力范围, 且 Phishpedia 受限于品牌库的更新时延。相比之下, DPFEAT 借助时空特征提取器能够实时解析页面结构, 并通过多分支特征融合与自适应加权, 在通用钓鱼检测场景中实现了更高的 Precision 与 Recall, 尤其擅长识别无品牌特征的钓鱼网站。

4.4.2 消融实验

为评估模型各组成部分的贡献, 本文进行了消融实验。实验设置了如下变体模型: DPFEAT-Full (完整模型)、DPFEAT-NoTempora (移除时序特征分支)、DPFEAT-NoSpatial (移除空间特征分支)、DPFEAT-NoFusion (移除自适应融合)、DPFEAT-NoAttention (移除注意力机制) 和 DPFEAT-SingleLoss (仅使用交叉熵损失)。消融实验结果对比见表 5。

实验结果表明, DPFEAT 模型的每个组成部分对性能提升均起到了重要的作用。DPFEAT-Full 在 Accuracy、Precision、Recall、F1-Score 和 AUC-ROC 指标上均表现最优。移除时序特征分支、空间特征分支、自适应融合模块或注意力机制均会导致模型的性能下降, 其中自适应融合模块对模型准确率的影响最大。这说明该模块能够动态调整各特征的权重, 提升模型对关键信息的关注。移除时序特征分支也带来了模型性能的下降, 这是因为钓鱼网站常模仿合法网站的 URL 结构, 其时序特征中蕴含的异常模式能够被时序特征分支有效捕捉。仅使用交叉熵损失的 DPFEAT-SingleLoss 性能下降相对较小, 但仍不及 DPFEAT-Full。

表 5 消融实验结果对比

变体模型	Accuracy	Precision	Recall	F1-Score	AUC-ROC
DPFEAT-Full	95.40%	96.30%	95.50%	95.00%	98.83%
DPFEAT-NoTempora	89.10%	88.30%	87.90%	88.10%	90.10%
DPFEAT-NoSpatial	88.50%	87.60%	87.20%	87.40%	89.40%
DPFEAT-NoFusion	86.70%	85.90%	85.40%	85.60%	87.80%
DPFEAT-NoAttention	87.90%	86.80%	86.50%	86.60%	88.70%
DPFEAT-SingleLoss	89.20%	88.10%	87.70%	87.90%	89.80%

4.5 损失函数对模型性能的影响

为评估损失函数对模型分类性能的影响，本文对分别采用 Focal+Dice、Focal 和 Dice 损失函数的模型开展了对比实验，结果见表6。

在 Accuracy 指标上，使用 Focal+Dice 损失函数的模型的表现（95.40%）优于单独使用 Focal 损失函数（89.10%）和单独使用 Dice 损失函数（93.40%）的模型。这说明 Focal+Dice 损失函数能够有效提升模型的整体预测能力。

在 Precision 指标上，Focal+Dice 组合模型（96.30%）同样优于单独使用 Focal 损失函数（95.30%）和单独使用 Dice 损失函数（91.80%）的模型，这说明 Focal+Dice 组合损失函数能够有效降低假正例的出现频率。

在 Recall 指标上，Focal+Dice 组合模型达到了 95.50%，而单独使用 Focal 损失函数和 Dice 损失函数的模型分别为 87.90% 和 94.20%。尽管单独使用 Dice 损失函数的模型在 Recall 上表现优秀，但 Focal+Dice 组合模型在保持高 Accuracy 的同时实现了优异的 Recall 表现，这使得其在处理类别不平衡数据集时具有更好的鲁棒性。

在 F1-Score 指标上，Focal+Dice 组合模型同样取得了最优表现（95.00%），这表明 Focal+Dice 组合模型在保持高 Precision 的同时，能够有效识别更多的正类样本。

在 AUC-ROC 指标上，单独使用 Focal 损失函数的模型达到了 97.10%，展现出优异的分类区分能力。而 Focal+Dice 组合模型也表现卓越，AUC-ROC 值达到 98.80%，证明该组合损失函数在类别判别任务中同样具有出色的表现。

4.6 可解释性分析框架

4.6.1 各模块对模型性能的影响

本文采用模块化设计思路，结合对各模块特征输出贡献度的分析，构建了一个全面可解释的分析体系。通过模块化设计能够直观地了解各模块对模型性能的影响，相关结果见表7。

表7 模块性能影响

模块	性能影响百分比	核心优势
自适应融合模块	9.12%	特征交互增强、自适应权重分配、多尺度信息融合
空间特征分支	7.23%	空间依赖建模、局部特征提取、空间关系捕获
时序特征分支	6.6%	时序模式识别、长期依赖建模、动态特征提取
注意力机制	7.86%	关键特征强化、注意力引导、特征选择优化
组合损失	6.49%	多目标优化、特征学习增强、模型泛化提升

为量化评估各模块（包括融合模块、空间特征分支、时序特征分支、注意力机制和组合损失函数）对模型整体性能的贡献，本文以完整模型（DPFEAT-Full）作为基准，采用逐步移除或替换特定组件的方法构建多个变体，观察和记录由此引起的性能变化。性能影响百分比的计算公式如下。

$$\text{性能影响百分比} = \frac{|\text{Baseline Metric} - \text{Variant Metric}|}{\text{Baseline Metric}} \times 100\% \quad (15)$$

其中，Baseline Metric 表示 DPFEAT-Full 在某项评估指标上的值，Variant Metric 表示变体在同一评估指标上的值。

观察表中数据可知，自适应融合模块、空间特征分支和注意力机制是提升模型性能的关键要素。自适应融合模块贡献了9.12%的性能提升，证明其有效增强了模型的特征交互和自适应能力；空间特征分支模块贡献了7.23%的性能，主要归因于其对空间特征的有效捕获，并与时序特征分支形成互补；注意力机制带来7.86%的性能提升，证明该机制通过精准的特征选择增强了模型的可解释性。而融合模块实现了6.6%的性能提升，证明该模块有效增强了模型的特征交互和自适应能力。这些模块协同作用，显著提高了模型的整体性能。

4.6.2 注意力权重对模型性能的影响

注意力权重是指模型在不同输入之间分配的

表6 损失函数对模型性能影响的对比结果

损失函数	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Focal+Dice	95.40%	96.30%	95.50%	95.00%	98.80%
Focal	89.10%	95.30%	87.90%	88.40%	97.10%
Dice	93.40%	91.80%	94.20%	93.40%	93.40%

权重值。这些权重决定了每个输入对模型最终预测的贡献程度。通过对注意力权重的可视化呈现,直观揭示模型决策时重点关注的输入部分,以及各输入之间的关联处理机制,进而深入解析模型的内在行为逻辑。为验证各特征权重差异的统计显著性,本文采用 Kruskal-Wallis 非参数检验($\alpha = 0.05$)对 SHAP 值的分布特征进行分析。结果显示,各特征权重的分布存在统计学显著差异(Kruskal-Wallis $H = 37.2, p < 0.001$)。进一步分析发现,“page_rank”和“google_index”等特征的权重显著高于其他特征($p < 0.01$),说明这两类特征在模型的判别过程中发挥关键作用。

图9展示了模型对各输入特征的重视程度。注意力权重值越大,表示该特征对模型预测结果的贡献越大。从图中可以看出,page_rank和google_index是模型最关注的两个特征,其权重值明显高于其他特征,这意味着page_rank和google_index可能是模型判断网页类别的重要依据。紧随其后的是nb_www和nb_hyperlinks,表明模型在判断网页类别时同样兼顾了这两个特征。此外,nb_hyphens、ratio_extHyperlinks和phish_hints等特征虽然权重较低,但对模型预测结果仍具有一定影响。此类可视化分析,有助于深入理解模型的决策逻辑,为后续模型的参数调整与优化提供依据。

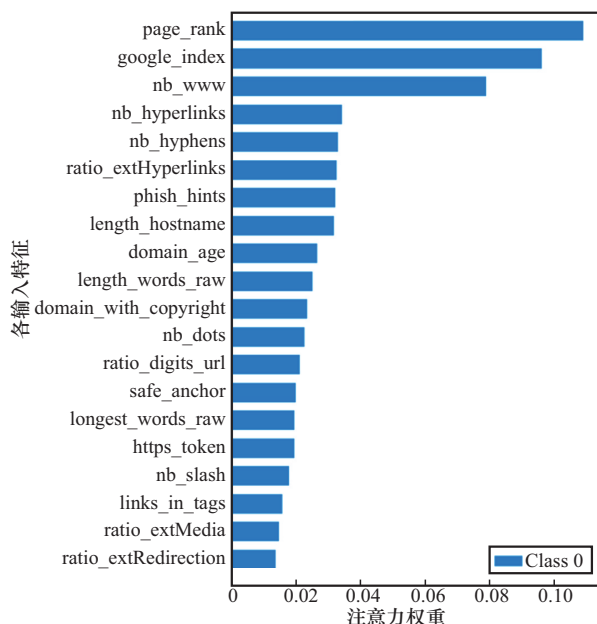


图9 各输入特征的注意力权重对比

4.7 计算成本和落地可行性

DPFEAT模型在NVIDIA 3090 GPU上训练耗时约2.2 h。其模块化设计不仅有效减少了对GPU计算资源的需求,节约了计算成本,也便于模型的部署与迭代升级。此外,该模型具有良好的识别效果和低误报率,有助于降低企业运维成本,可集成至安全网关或在线钓鱼监控系统。

5 结束语

本文提出的DPFEAT模型在钓鱼网站检测任务中取得了显著进展,通过创新的架构设计和优化策略,实现了性能的全面提升。所构建的可解释性分析框架为模型决策过程提供了透明的解释。实验结果表明,该模型在准确率、精确率等关键指标上均优于现有方法,展现出较高的实用价值。其优异的性能与泛化能力,使其具有良好的应用前景。未来的研究工作将聚焦于模型轻量化、算法效率提升和应用场景拓展,以进一步提升模型的实用价值。

参考文献:

- [1] 朱世起, 努尔布力. 钓鱼网站检测研究现状与发展趋势的计量分析[J]. 计算机工程与应用, 2020, 56(15): 92-100.
Zhu S Q, Nurbol. Bibliometric analysis of current studies and developing trends on phishing sites detection[J]. Computer Engineering and Applications, 2020, 56(15): 92-100.
- [2] 蔡召, 荆涛, 任爽. 以太坊钓鱼诈骗检测技术综述[J]. 网络与信息安全学报, 2023, 9(2): 21-32.
Cai Z, Jing T, Ren S. Survey on Ethereum phishing detection technology[J]. Chinese Journal of Network and Information Security, 2023, 9(2): 21-32.
- [3] 王雨琪, 刘博文, 林果园. 基于URL语言特征的钓鱼网站检测算法[J]. 计算机工程与应用, 2019, 55(24): 84-90.
Wang Y Q, Liu B W, Lin G Y. Phishing detection algorithm based on language features of URL[J]. Computer Engineering and Applications, 2019, 55(24): 84-90.
- [4] 顾晓清, 王洪元, 倪彤光, 等. 基于贝叶斯和支持向量机的钓鱼网站检测方法[J]. 计算机工程与应用, 2015, 51(4): 87-90, 95.
Gu X Q, Wang H Y, Ni T G, et al. Phishing detection approach based on Bayes and support vector machine[J]. Computer Engineering and Applications, 2015, 51(4): 87-90, 95.
- [5] Zhou J X, Cui H B, Li X N, et al. A novel phishing website detection model based on LightGBM and domain name features[J]. Symmetry, 2023, 15(1): 180.

- [6] Huang Y. Phishing detection using convolutional neural networks and feature engineering[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(10): 2547-2558.
- [7] Wang L, Liu Y. Phishing detection using LSTM and feature engineering[J]. IEEE Transactions on Information Forensics and Security, 2020, 15(9): 2041-2050.
- [8] 文伟平, 朱一帆, 吕子晗, 等. 针对品牌的网络钓鱼扩线与检测方案[J]. 信息安全学报, 2023, 23(12): 1-9.
Wen W P, Zhu Y F, Lyu Z H, et al. Brand-specific phishing expansion and detection solutions[J]. Netinfo Security, 2023, 23(12): 1-9.
- [9] 弋晓洋, 张健. 基于图像的网络钓鱼邮件检测方法研究[J]. 信息安全学报, 2021, 21(9): 52-58.
Yi X Y, Zhang J. Image-based phishing email detection method and implementation[J]. Netinfo Security, 2021, 21(9): 52-58.
- [10] 张桥, 卜佑军, 陈博, 等. 一种基于MPAN的钓鱼URL检测方法[J]. 信息工程大学学报, 2021, 22(4): 443-449.
Zhang Q, Bu Y J, Chen B, et al. Phishing URL detection method based on MPAN[J]. Journal of Information Engineering University, 2021, 22(4): 443-449.
- [11] Ma C Y, Wu A M, Ma W Z, et al. Malicious URL recognition based on multi-feature fusion and machine learning[C]//Proceedings of the 2022 41st Chinese Control Conference (CCC). Piscataway: IEEE Press, 2022: 3014-3019.
- [12] 张思睿, 延志伟, 董科军, 等. 基于注意力机制与特征融合的钓鱼网站检测算法[J]. 网络与信息安全学报, 2024, 10(4): 123-131.
Zhang S R, Yan Z W, Dong K J, et al. Phishing detection algorithm based on attention and feature fusion[J]. Chinese Journal of Network and Information Security, 2024, 10(4): 123-131.
- [13] Bahaghghat M, Ghasemi M, Ozen F. A high-accuracy phishing website detection method based on machine learning[J]. Journal of Information Security and Applications, 2023, 77: 103553.
- [14] Tan C C L, Chiew K L, Yong K S C, et al. Hybrid phishing detection using joint visual and textual identity[J]. Expert Systems with Applications, 2023, 220: 119723.
- [15] Johnson O V, Chen X Y, Khaw K W, et al. Ps-CALR: periodic-shift cosine annealing learning rate for deep neural networks[J]. IEEE Access, 2023, 11: 139171-139186.
- [16] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2017: 2999-3007.
- [17] Liu Q, Liu H T, Zhao Y, et al. Dual-branch network with dual-sampling modulated dice loss for hard exudate segmentation in color fundus images[J]. IEEE Journal of Biomedical and Health Informatics, 2022, 26(3): 1091-1102.
- [18] Adebowale M A, Lwin K T, Hossain M A. Intelligent phishing detection scheme using deep learning algorithms[J]. Journal of Enterprise Information Management, 2023, 36(3): 747-766.
- [19] Aljofey A, Jiang Q S, Qu Q, et al. An effective phishing detection model based on character level convolutional neural network from URL[J]. Electronics, 2020, 9(9): 1514.
- [20] Zhu Y M, Wang M, Yin X F, et al. Deep learning in diverse intelligent sensor based systems[J]. Sensors, 2023, 23(1): 62.
- [21] Benavides-Astudillo E, Fuertes W, Sanchez-Gordon S, et al. Comparative study of deep learning algorithms in the detection of phishing attacks based on HTML and text obtained from web pages[C]//Applied Technologies. Cham: Springer Nature Switzerland, 2023: Proceedings of the 4th International Conference on 386-398.
- [22] Ponni P, Prabha D. Randomized active learning to identify phishing URL[M]//Advanced Communication and Intelligent Systems. Cham: Springer Nature Switzerland, 2023: 533-539.
- [23] Mijwil M M, Salem I. E, Ismael M M. The significance of machine learning and deep learning techniques in cybersecurity: a comprehensive review[J]. Iraqi Journal for Computer Science and Mathematics, 2023: 87-101.
- [24] Ghosh M, Srivastava S, Upadhyaya A, et al. CATALOG: exploiting joint temporal dependencies for enhanced phishing detection on ethereum[C]//Proceedings of the ACM on Web Conference 2025. New York: ACM, 2025: 969-977.
- [25] Lin Y, Liu R F, Divakaran D M, et al. Phishpedia: a hybrid deep learning based approach to visually identify phishing webpages[C]//Proceedings of the 30th USENIX Security Symposium. Berkeley: USENIX Association, 2021: 3793-3810.

[作者简介]



郭凯 (1999-), 男, 山西吕梁人, 宁夏大学硕士生, 主要研究方向为网络空间安全。



王恒 (1975-), 男, 宁夏银川人, 博士, 宁夏大学教授, 主要研究方向为网络空间安全、自然语言处理。



马自强 (1990-), 男, 新疆乌鲁木齐人, 博士, 宁夏大学副教授, 主要研究方向为密码应用、系统安全、区块链、恶意流量识别、网络舆情分析。