基于深度残差神经网络的博彩网页识别算法设计

张　聪，张　恒，张立坤，赵　彤，邓桂英
（中国互联网络信息中心 技术研发部，北京 100190）

摘　要： 互联网对人民群众的生活和工作产生了重要影响，然而网络空间中隐藏着大量有害的博彩网站或赌博网站，很容易给网民造成损失和困扰，甚至可能扰乱社会秩序，因而研究对此类网站进行高效识别的方法具有重要意义。提出利用深度残差神经网络解决博彩类网页识别问题，基于深度残差网络的原理设计了算法 GamblingRec。经验证，算法准确率达到了 95.16%，正样本召回率为 93.21%，表明基于深度残差神经网络的方法能够用于博彩类网页识别，并能达到较高的识别性能。

关键词： 卷积神经网络；残差网络；博彩；网页分类；ResNet
中图分类号： TN91　　**文献标识码：** A　　**DOI：** 10.16157/j.issn.0258-7998.211757

中文引用格式： 张聪，张恒，张立坤，等. 基于深度残差神经网络的博彩网页识别算法设计[J].电子技术应用，2022，48(2)：15-18.
英文引用格式： Zhang Cong，Zhang Heng，Zhang Likun，et al. Gambling web page recognition algorithm design based on deep residual neural network[J]. Application of Electronic Technique，2022，48(2)：15-18.

Gambling web page recognition algorithm design based on deep residual neural network

Zhang Cong，Zhang Heng，Zhang Likun，Zhao Tong，Deng Guiying
(Technological Research and Development Department，China Internet Network Information Center(CNNIC)，Beijing 100190，China)

Abstract： The Internet has an important impact on people's life and work. However，there are a large number of harmful gambling websites hidden in cyberspace，which is easy to cause losses and troubles to netizens，it can even disturb society order. Therefore，it is of great significance to study the efficient recognition method of such websites. In this paper，the deep residual neural network is used to solve the problem of gambling web page recognition，and the algorithm GamblingRec is designed based on principle of deep residual network. The results show that the accuracy of GamblingRec reaches 95.16%，and the positive sample recall rate is 93.21%，which indicates that the method based on deep residual neural network can be applied for gambling web page recognition，and can achieve high recognition performance.

Key words： convolutional neural network；residual network；gambling；web page classification；ResNet

0 引言

随着互联网技术的高速发展，我国网民人数持续增长，根据《第 47 次中国互联网络发展状况统计报告》的数据，截至 2020 年 12 月，我国网民人数已达到 9.89 亿[1]，毫无疑问，互联网已经成为人们日常生活不可或缺的一部分。然而，虚拟的网络空间中隐藏着大量有害的博彩类型网站，极易给参与者造成经济损失，设计有效方法对博彩类网站进行识别具有重要意义。

1 相关工作

博彩网站识别相当于对网页进行分类，预测其为博彩网页或其他类型网页。付顺顺[2]采用 FastText[3]算法和 Bootstrap[4]集成算法，利用网站文本数据，提高了识别速度并减轻了正常网站和博彩网站数据不均衡问题。唐喆[5]等人采用 SVM[6]算法并提取不同的文本特征，实现对网页的分类。

已有的网页识别方法常利用网页的结构化数据，人工构造基于规则的特征，然后结合机器学习等技术进行预测和识别。很少直接采用非结构化的网页图像作为模型的输入和训练数据，导致算法无法利用网页中十分重要的图形图像信息。

近些年来，随着神经网络算法的进步和硬件算力的提升，卷积神经网络(Convolutional Neural Network，CNN)[7-8]在图像分类和识别中取得了显著的研究成果。直接将网页图像作为模型的输入，利用深度神经网络的强大特征提取能力对网页特征进行提取，进而预测网页类型很有研究和工程应用价值。本文利用现代残差神经网络技术

设计了 GamblingRec，实现了对博彩类网页图像的自动特征提取并预测和识别网页类别。

2 设计方法

2.1 博彩网页图形特点

博彩类网站与其他类型网站相比通常有比较明显的区别，一般会将赌博的游戏载体进行图像化、卡通画，从而可以吸引眼球，这些载体有卡通鱼、棋牌、球类等，如图 1 所示，常用的关键词汇也常采用艺术化的形式展示。而 CNN 算法，具有自动提取图像特征[7]的能力，很适合应用于此类图形特点明显的场景，本文基于残差神经网络[9]原理设计了算法，实现对该类网页的识别。

图 1　博彩网页图像特征示例

2.2 残差神经网络

基本残差块由残差函数、跳跃连接和输出激活函数组成，通过跳跃连接将输入 X 与残差函数 $F(X)$ 输出相加，W、H、C 表示图像特征图的维度，求和结果经过激活函数 $G(X)$ 作为残差模块的最后输出。如图 2 所示，其中 $H(X)=F(X)+X$，输出 $Y=G(H(X))$。图 3 在恒等映射上增加了尺度变换，是为了保持恒等映射尺寸与残差函数输出尺寸保持一致。

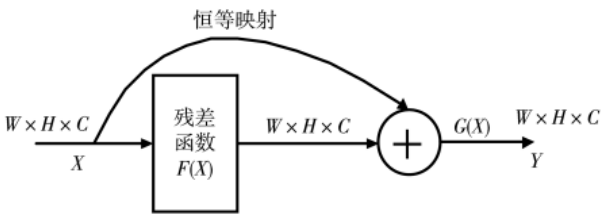

图 2　残差块 1

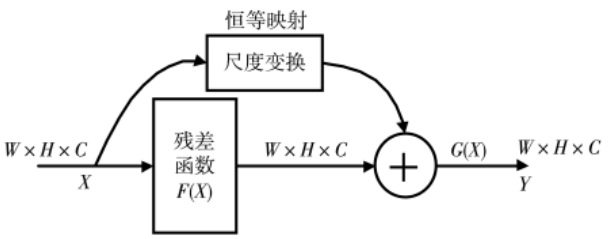

图 3　残差块 2

残差网络有助于解决深度网络的梯度消失和梯度爆炸问题，假设 X_l 和 X_{l+1} 分别表示第 l 个残差块的输入和输出；$F(X)$ 为残差函数，表示残差网络学习到的残差；公式中的 W_l 和 W_i 为残差函数 $F(X)$ 的可学习参数；G 表示 ReLU 激活函数[10]，则残差模块可以表示为：

$$x_{l+1}=x_l+F(x_l,W_l) \tag{1}$$

给定第 l 个残差模块输入 x_l，通过递归，可以求得第 L 个残差模块的输出为：

$$x_L=x_l+\sum_{i=l}^{L-1}F(x_i,W_i) \tag{2}$$

假设残差网络的损失函数为 $\mathrm{Loss}(w)$，w 表示网络参数，根据神经网络的链式求导法则，可得：

$$\frac{\partial \mathrm{Loss}}{\partial x_l}=\frac{\partial \mathrm{Loss}}{\partial x_L}\frac{\partial x_L}{\partial x_l}=\frac{\partial \mathrm{Loss}}{\partial x_L}\left[1+\frac{\partial}{\partial x_L}\sum_{i=l}^{L-1}F(x_i,W_i)\right] \tag{3}$$

从式(3)可知，即使是很深的网络层，输出误差也可以无损地传播到网络最初的输入网络层(第 l 层)，从而避免了普通深层网络长传播路径所带来的梯度消失或爆炸的问题。

2.3 模型设计

深度残差网络一般通过多组残差块的堆叠达到较强的特征提取能力，并可以获得良好的性能。本文设计了包含 9 个卷积层和 1 个全连接层的深度残差网络 GamblingRec，具体的设计参数如表 1 所示。网络结构如图 4 所示，主要由卷积层(Conv)、池化层(Pooling)、Dropout 层和全连接层(FC)构成[11-13]，Conv1 的卷积核尺寸为 7×7，卷积核个数有 64 个，卷积步长(stride)为 2。Conv2~5 采用图 3 所示的残差块结构，残差函数包含两个卷积层，且第一个卷积层进行步长为 2 的卷积运算，恒等映射路径上通过步长为 2 的 1×1 卷积运算将恒等映射输出的深度和宽度转化为与残差函数的输出特征图尺寸一致。在经过 4 个卷积块处理后连接一个 Dropout 层，以概率 0.2 随机丢弃神经元来提高模型的泛化能力，然后通过全局平均池化提取每个通道特征图的特征，最后使用全连接 FC 层和 softmax[14]函数输出两种类别的预测值。

表 1　模型架构

层	参数
Conv1	7×7，64，stride 2
Pooling	3×3 max pooling(最大池化)，stride 2
Conv2	$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times1$，stride 2
Conv3	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times1$，stride 2
Conv4	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times1$，stride 2
Conv5	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times1$，stride 2
Dropout	Dropout，p=0.2
Pooling-avg	Average pooling(平均池化)
FC	2-d，softmax

3 图像数据扩增

卷积神经网络模型是一种数据驱动的算法，拥有大

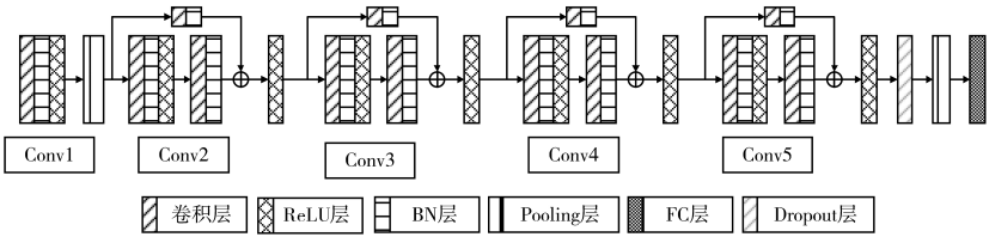

图 4　GamblingRec 网络架构

规模数据对于模型的训练十分重要，更多的数据能够提升模型的泛化能力，但获取大量的真实数据有时候并非那么容易，这时就要考虑采用数据增强技术[15]。数据增强不用实际收集新数据，却可以达到增加数据的效果，它在原始数据的基础上进行数据变换，增加了数据多样性，有助于降低模型过拟合并提升模型准确率。

数据增强对于图像数据而言更加有效，例如两张存在一定位移的图像，在人看来可能没什么区别，但在神经网络看来却是截然不同的数据，图像数据增强极大地增加了神经网络所能看到的图像多样性。

针对博彩网页图像数据，本文进行了随机水平翻转、随机旋转、随机灰度变化、随机截取的图像变换技术，最后对输入图像数据进行标准化处理，提升模型训练的速度和稳定性，如图 5 所示。

图 5　图像数据增强流程

4 实验介绍

4.1 实验数据

本文实验用的数据集分为训练集和验证集，为增加数据量并对图像尺寸进行统一化，将原始网页图像裁剪为高和宽都为 1 000 的正方形，实际训练过程中还采用数据扩增技术，详情见第 3 节，利用图像的随机变换大大增加了训练样本的数量，这有利于增加模型的泛化水平。经过图像扩增后的图像尺寸为 600×600，训练集样本数量为 15 649，验证集样本数量为 3 509，正样本与负样本的比例为 7.3:10，硬件采用 GPU 进行训练，型号为 Tesla P100。

4.2 实验设置

模型优化采用交叉熵[14]作为损失函数，具体实现使用了 PyTorch 提供的交叉熵损失函数 CrossEntropyLoss。卷积神经网络训练所采用的超参数如表 2 所示，训练时通过随机的图像变换，图像尺寸转换为 600×600，优化器采用随机梯度下降[16]，学习率衰减采用 StepLR，初始学习率为 0.01，每隔 5 个 Epoch 将学习率下降 10%。

4.3 实验结果分析

4.3.1 评估指标

本文采用 4 个指标评估模型在验证集上的表现，分别是召回率(Recall)、精确率(Precision)、准确率(ACC)和 $F1$ 分值。TP 表示被正确分类的不良图片数量；FN 表示被误判为良性图片的不良图片数量；$F1$ 分值是 Recall 和 Precision 的调和平均，反映了 Recall 和 Precision 的整体表现，一般 $F1$ 越大模型表现就越好；ACC 为验证集上整体准确率。如下所示为这几种指标的定义：

$$\mathrm{ACC}=\frac{\mathrm{TP}+\mathrm{TN}}{\text{样本总数}} \tag{4}$$

$$\mathrm{Recall}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{5}$$

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

$$F1=\frac{2\mathrm{Recall}\cdot\mathrm{Precision}}{\mathrm{Recall}+\mathrm{Precision}} \tag{7}$$

表 2　模型训练超参数

参数	大小
Epoch	200
图像尺寸	600×600×3
优化器	SGD
学习率衰减策略	StepLR，step=5
初始学习率/衰减率	0.01/0.9
Moment(动量)	0.9
Weight_decay(权值衰减)	1×10^{-4}
Batch size(每批数量)	64

4.3.2 实验结果

实验表明采用深度残差网络输入网页截图可以有效识别出网页类别，且具有较高的准确率。表 3 列出了模型在验证集上的不同评测指标，主要为正样本的精确率和召回率、负样本的精确率和召回率，以及验证集全样本的准确率。从表中可以看出，模型准确率达到了 95.16%，正样本精确率和召回率分别为 96.01%和 93.21%，正样本的召回率略低于负样本的召回率。图 6 的准确率(ACC)曲线展示了模型训练过程中准确率的变化情况，从图中可以看出训练集的准确率稳步提升，而验证集的准确率先是经过了 100 多个 Epoch 的震荡，然后逐渐稳

表 3　验证集评测结果　(%)

	Precision	Recall	ACC
正样本	96.01	93.21	95.16
负样本	94.48	96.78	

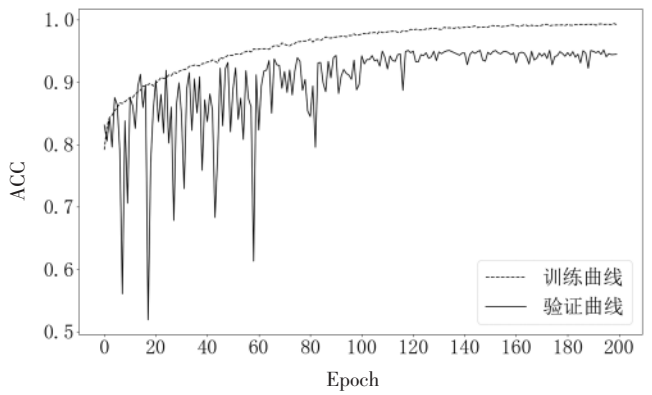

图 6 准确率曲线

定在 95%左右。图 7 为损失函数(Loss)曲线,也是先经过 100 多个 Epoch 的震荡并逐渐收敛。

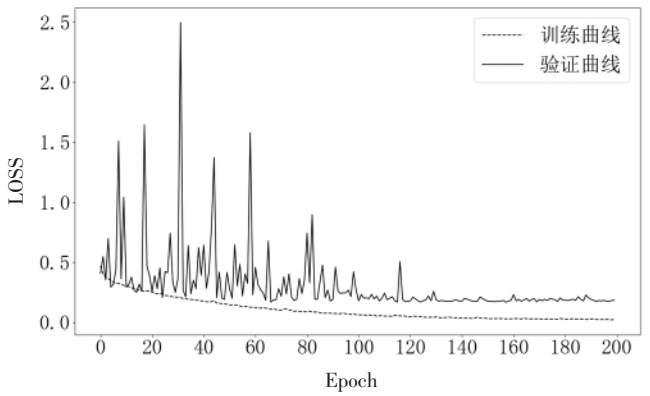

图 7 损失函数曲线

5 结论

本文针对博彩类网站的识别问题,从网络爬取了博彩类网页数据和其他类型网页数据,构建了博彩图像数据集。基于深度残差网络的方法设计了 GamblingRec,进行了模型训练和优化,并在验证集上对博彩网页进行识别和评测,获得了良好的识别效果。但由于网页形式和内容多种多样,除了图形还有文字,有的网页图形特征明显,而有的网页以文字为主,未来有必要研究能同时兼顾图形和文字的识别方法。另外,进一步收集更多的数据,使模型对网页形式和内容有更强的适应性,还需要研究提升正样本召回率的方法。

参考文献

[1] CNNIC.第 47 次中国互联网络发展状况统计报告[R].北京:中国互联网络信息中心,2021.
[2] 付顺顺.基于 FastText 的赌博网站识别方法[J].网络安全技术与应用,2020(8):150-151.
[3] JOULIN A,GRAVE E,BOJANOWSKI P,et al.Bag of tricks for efficient text classification[C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics,2017,2:427-431.
[4] EFRON B.Bootstrap methods:another look at the jackknife[J].The Annals of Statistics,1979,7(1):1-26.
[5] 唐喆,曹旭东.网页分类中特征选择方法的研究[J].电子设计工程,2016,24(5):120-122.
[6] SMOLA A J,SCHÖLKOPF B.A tutorial on support vector regression[J].Statistics and Computing,2004,14(3):199-222.
[7] YAMASHITA R,NISHIO M,DO R K G,et al.Convolutional neural networks:an overview and application in radiology[J].Insights Imaging,2018(9):611-629.
[8] Gua Jiuxiang,Wang Zhenhua,KUEN J,et al.Recent advances in convolutional neural networks[J].Pattern Recognition:The Journal of the Pattern Recognition Society,2018(77):354-377.
[9] He Kaiming,Zhang Xiangyu,Ren Shaoqing,et al.Deep residual learning for image recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition,2016:770-778.
[10] GLOROT X,BORDES A,BENGIO Y.Deep sparse rectifier neural networks[J].Journal of Machine Learning Research,2011(15):315-323.
[11] SRIVASTAVA N,HINTON G,KRIZHEVSKY A,et al.A simple way to prevent neural networks from overfitting[J].Journal of Machine Learning Research,2014,15(1):1929-1958.
[12] O'SHEA K,NASH R.An introduction to convolutional neural networks[J].arXiv preprint arXiv:1511.08458,2015.
[13] LECUN Y,BENGIO Y.Convolutional networks for images,speech,and time series.The handbook of brain theory and neural networks[M].MIT Press,1995.
[14] 邓建国,张素兰,张继福,等.监督学习中的损失函数及应用研究[J].大数据,2020,6(1):60-80.
[15] KRIZHEVSKY A,SUTSKEVER I,HINTON G E.ImageNet classification with deep convolutional neural networks[J].Communications of the ACM,2017,60(6):84-90.
[16] 史加荣,王丹,尚凡华,等.随机梯度下降算法研究进展[J].自动化学报,2021,47(9):2103-2119.

(收稿日期:2021-05-14)

作者简介:

张聪(1987-),男,硕士,工程师,主要研究方向:人工智能、数据科学、电子系统设计。

张恒(1978-),男,硕士,工程师,主要研究方向:机器学习、数据分析。

张立坤(1983-),男,硕士,高级工程师,主要研究方向:大数据、机器学习。

扫码下载电子文档

版权声明