

基于 OCR 技术的网页篡改检测研究^①

吴伟斌

(泉州师范学院网络中心 福建泉州 362000)

摘要 [目的] 当前网站受到各种安全的威胁态势并未减弱, 网站被篡改事件还在增长, 除了加强防范外, 网页被篡改检测也是重要一环。[方法] 本文提出基于 OCR (Optical Character Recognition) 技术的网页篡改检测模型, 利用自然场景文字识别技术对网页截图进行文字识别, 经敏感词检测判断网页是否被篡改。[结果] 经对网页篡改检测模型进行实验, 该模型在实验的数据集上准确率较高。[局限] 模型也存在不足之处: 无法识别隐式篡改; 依赖敏感词, 文字之外其他异常信息尚待研究。[结论] 基于 OCR 技术的可用于网页文字篡改的检测。

关键词 OCR; 网页; 篡改检测; 场景文字识别

Research on Web Page Tampering Detection Based on OCR Technology

Wu Weibin

(Network Center of Quanzhou Normal University, Quanzhou, Fujian, 362000, China)

Abstract [Objective] At present, the threat situation of website security has not weakened, and the number of website tampering incidents is still increasing. In addition to strengthening prevention, the detection of webpage tampering is also an important part. [Methods] This paper proposes a web page tamper detection model based on OCR (Optical Character Recognition) technology. It uses scene text recognition technology to recognize the characters of web page screenshots, and judges whether the web page has been tampered by sensitive word detection. [Results] Through experiments on the web page tampering detection model, the model has high accuracy on the experimental data set. [Limitations] There are also deficiencies in the model: it is impossible to identify implicit tampering; relying on sensitive words and other abnormal information outside the text remains to be studied. [Conclusions] the method based on OCR technology can be used to detect the tampering of web pages.

Keywords OCR; Webpage; Tampering Detection; Scene Text Recognition

①本文系 2018 年福建省中青年教育科研项目 (信息化专项) (项目编号: JZ180189) 的研究成果之一。

1 引言

在信息时代,互联网已成为生活的一部分,网站成为人们获取信息的重要来源,已成为信息传播、电子商务、电子政务等的重要载体。虽然现在网站已采取不少的安全防范措施,但由于系统漏洞问题会长期存在,病毒、木马和恶意代码肆虐,网站的被植入后门、被篡改等事件态势并未减弱。网站被篡改事件迅猛增长,已经成为危害比较严重的网络安全问题,截至2019年12月,国家计算机网络应急技术处理协调中心监测发现我国境内被篡改网站185573个,较2018年底(7049个)增长较多^[1]。

网络安全问题严重影响网站建设单位的形象,可能造成一定的经济损失和不良社会影响,需构建一个完善的网络安全体系。网站的前置各种软硬件防火墙、Web应用防护系统、入侵检测系统等安全产品也在不断发展进步,网站系统本身也建立了各种防篡改的防护,但篡改事件仍持续发生。为减少网站被篡改所带来的影响,除了建立有效的防止网页篡改措施,还需继续研究如何在网站被篡改后及时检测识别等工作。

2 相关工作

常见的防网页篡改技术主要有外挂轮巡技术、核心内嵌技术、事件触发技术^[2]。三种方法中核心内嵌技术、事件触发技术适用于网络管理员在操作系统、服务器上主动部署、主动防御;而外挂轮巡技术即可应用于服务器侧,也可以应用于第三方检测^[3]。外挂轮巡技术主要是网页与其基线进行比较来判断完整性,但其对动态网页进行检测、其对网页图片篡改类的识别效果很差,有一定的局限性。为弥补该技术的短板,引进图像识别技术

来解决此类问题^[4]。文献[3]是用角点检测技术用已知样本图片对网页篡改截图进行有效识别标记。文献[4]则是利用图像处理中两种特征点检测的方式获取图像中的特征点信息从而起到检测网页是否被篡改的作用。

网页篡改按照攻击手段来分类,有显式篡改和隐式篡改两种方式^[5]。本文主要研究显式篡改检测识别,通过网页截图对网页篡改检测进行识别,利用OCR文字识别后综合判断检测网页是否被篡改。不少人将OCR技术定义为广义的所有图像文字检测和识别技术(简称图文识别技术),即包括传统的OCR识别技术,又包括自然场景文字识别技术^[6-7],本文所用OCR指的是广义OCR技术,将利用自然场景文字识别技术。PaddleOCR旨在打造一套丰富、领先且实用的OCR工具库^[8],供多种文字检测训练算法和多种文字识别训练算法,本文使用PaddleOCR提供的中文OCR模型完成文本检测以及文本识别串联任务,对网页截图进行文字识别,对识别后的文本进行敏感词检测,判断网页是否被篡改。

3 篡改检测模型

3.1 篡改检测模型实现

具体检测模型实现步骤如下:①模拟访问待检测网站,抓取网页并截图保存;②OCR识别截图中的文字生成文本;③对生成的文本进行敏感词检测。后续若是检测被篡改发送报警到管理员,非本文重点讨论不再赘述。网页的篡改检测流程如图1所示。

3.2 OCR文字识别

自然场景文字识别技术主要包括两个步骤:文本检测和文本识别,流程如图2所示。文字检测上找出文字位置和范围;文本识别对文本检测定位的文本区域进行识别,输出结果:文本信息。



图1 网页的篡改检测流程

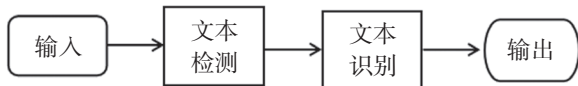


图2 文字识别技术流程

文字识别是将图片中的文字序列识别的过程。文字识别时输入的是含有文字的候选框，输出是该检测框中的文字序列^[9]。PaddleOCR在文本检测和文本识别分别提供多种算法，下文实验文本检测采用文本检测模型DB（Differentiable Binarization）^[10]算法，文本识别采用CRNN（Convolutional Recurrent Neural Network）^[11]算法。

DB是一个检测任意形状的场景文本检测网络，利用可微分二值化的方法，可在分割网络自适应地设定二值化阈值。可微分二值化不仅有助于把文字区域与背景区分开，而且还能把相近的实例分离开来^[12]。在ICDAR2015文本检测公开数据集上，算法效果精度83.79%^[8]。

CRNN网络是一个端到端的识别网络，包括特征提取、序列分析、序列解码三个部分。该网络首先利用CNN（Convolutional Neural Networks）提取文本图像的特征，使用双向LSTM（Long Short-Term Memory）^[13]提取上下文特征得到每列特征的概率分布，将语音识别领域的CTC（Connectionist Temporal Classification）^[14]引入图像处理不定长序列对齐问题，对输入的单个词的切分和序列进行整合。DB和CRNN两个模型串联使用，分别实现文本检测和文本识别。

4 实验与分析

通过工具采集了正常99个网站首页的截图

及网上收集篡改网页（包含敏感词）截图11张图片；对正常99个网页的截图模拟网页被篡改操作，通过图片处理技术加入敏感词，然后对图片进行随机仿射、加噪点等处理，共计产生内含敏感词的图片246张；总计获得实验数据356张图片，其中正常网页图片99张，被篡改网页图片257张。

实验硬件环境为虚拟机：Xeon CPU E5-2650 v4、内存48G，软件环境CentOS 8、python 3.7。实验采用PaddleOCR作为文字识别模型，使用通用中文OCR模型（简称为DB_CRNN）和支持空格的通用中文OCR模型（简称为DB_CRNN_EN）^[8]两个模型进行OCR文字识别，其中文本检测采用DB算法、文本识别采用CRNN算法。这两个模型分别对图片进行OCR文字识别并生成对应的文本，对生成文本进行敏感词检测。

结果评价采用准确率和二值分类器常见的两个指标假正率（False Positive Rate, FPR）、假负率（False Negative Rate, FNR）。FPR和FNR在这里定义如下。

$$FPR = \frac{\text{被认为正常的篡改页面数}}{\text{篡改页面总数}}$$

$$FNR = \frac{\text{被认为异常的正常页面数}}{\text{正常页面总数}}$$
^[15]

两个模型的运行时间只计算OCR文字识别的运行时间，敏感词检测时间短，忽略不计。

如表1所示,通过OCR识别网页截图的篡改检测结果, DB_CRNN_EN模型虽然运行平均时间比DB_CRNN模型略长,但假正率较低,检测准确率都高;假负率都为零,说明OCR识别网页截图的文本出现敏感词的异常概率接近零,也要选择好敏感词以免正常网页误判为异常。

网页截图进行OCR文字识别后的文本进行敏感词检测为最后步骤。如图3所示,横坐标为敏感词代码,共计13个词,纵坐标为检测出

敏感词的总数。被篡改网页的图片都包括有两个敏感词,其代码为分别S01和S02,图3显示两个模型都未能全部检测出来,经检查主要原因还是图片在加入敏感词的处理时包含两个敏感词的文字都不全,但配合其他敏感词判断准确率接近100%。同样的数据,对两个模型检测出敏感词的总数统计显示DB_CRNN_EN模型较优。

表 1 两个模型的 FPR、FNR 和平均值运行时间

模型	准确率/%	FPR/%	FNR/%	平均运行时间/s
DB_CRNN	99.16	1.17	0	48.06
DB_CRNN_EN	99.72	0.39	0	48.76

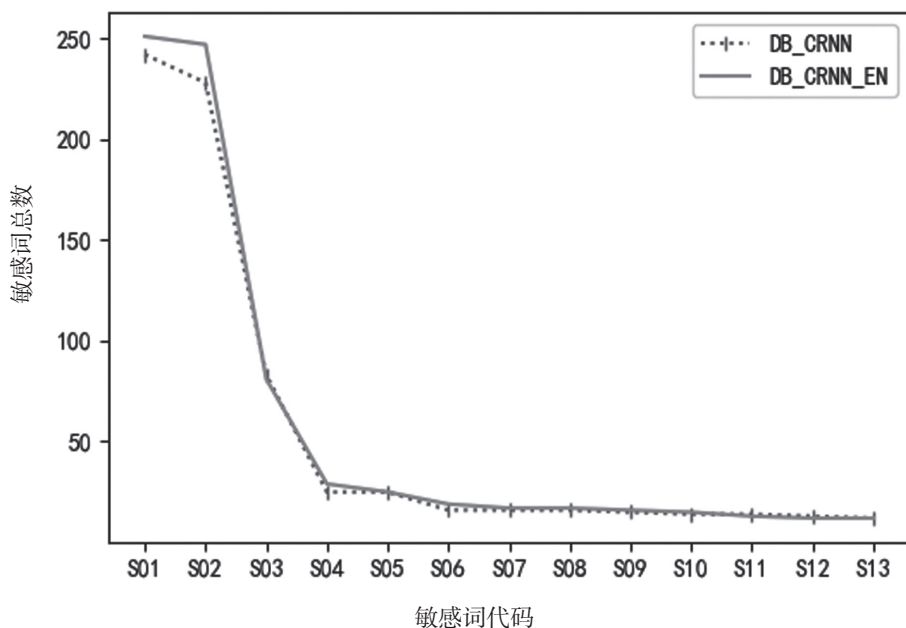


图 3 两种模型敏感词检测结果对比

5 结语

OCR技术的发展,特别是自然场景文字识别技术快速迭代,为网页截图文字识别提供了可能。本文提出利用OCR技术的网页篡改检测模型,采用PaddleOCR框架和文本检测DB

算法与文本识别CRNN算法对网页截图进行文字识字,对DB_CRNN和DB_CRNN_EN模型进行实验,两个模型都有较优的效果、准确率高,基于OCR技术的模型可用于网页文字篡改的检测。该模型也存在不足之处:无法识别隐

式篡改; 依赖敏感词, 文字之外其他异常信息的检测尚待研究。

参考文献

- [1] 中国互联网络信息中心. 第45次中国互联网络发展状况统计报告 (2020-04-28) [EB/OL]. [2020-08-08]. http://www.cac.gov.cn/2020-04/27/c_1589535470378587.htm.
- [2] 盖玲. 防网页篡改技术比较分析 [J]. 图书与情报. 2007 (01): 92—94.
- [3] 韩钢. 国家互联网应急中心图像处理技术在网页篡改识别上的应用 [J]. 通信管理与技术. 2017 (03): 57—58+61.
- [4] 颜于凤, 沈勇. 基于图像处理的网页篡改检测 [J]. 计算机与数字工程. 2020 (6): 1479—1482, 1518.
- [5] 王闻祎. 网页篡改检测系统设计与实现 [D]. 四川: 西南交通大学, 2019.
- [6] 牛小明, 毕可骏, 唐军. 图文识别技术综述 [J]. 中国体视学与图像分析. 2019 (24): 241—256.
- [7] SIGAI. 自然场景文本检测识别技术综述 (2018-06-30) [EB/OL]. [2020-08-08]. <https://blog.csdn.net/SIGAICSDN/article/details/80858565>.
- [8] PADDLEPADDLE. PaddleOCR (2020-08-08) [EB/OL]. [2020-08-08]. <https://github.com/PaddlePaddle/PaddleOCR>.
- [9] 白翔, 杨明锟, 石葆光, 等. 基于深度学习的场景文字检测与识别 [J]. 中国科学: 信息科学, 2018, 48 (5): 531—544.
- [10] LIAO M, WAN Z Y, YAO C, et al. Real-time Scene Text Detection with Differentiable Binarization [C]. National Conference on Artificial Intelligence, AAAI 2020: 11474—11481.
- [11] SHI B, BAI X, YAO C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition [J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39 (11): 2298—2304.
- [12] 墨殇浅尘. 场景文本检测 (Differentiable Binarization) -DB (2020-07-03) [EB/OL]. [2020-08-08]. <https://www.cnblogs.com/monologuesmw/p/13223314.html>.
- [13] GRAVES A, SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures [J]. Neural Networks, 2005 (5): 602—610.
- [14] GRAVES A, FERNÁNDEZ S, GOMEZ F, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks [C]. Proceedings of the 23rd international conference on Machine learning. New York: ACM, 2006: 369—376.
- [15] 魏文哈, 邓一贵. 基于局部变化性的网页篡改识别模型及方法 [J]. 计算机应用. 2013, 33 (2): 430—433.