计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(10):121−126 [doi: 10.15888/j.cnki.csa.006604]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

面向网站图像数据的安全分析系统①

王赛赛[1,2], 张　磊[1], 李　健[1]

[1](中国科学院 计算机网络信息中心, 北京 100190)
[2](中国科学院大学 计算机与控制学院, 北京 100049)
通讯作者: 王赛赛, E-mail: wangsaisai@cnic.cn

摘　要: 随着网络大众媒体的出现与流行, 产生了文字、图像、视频等各种类型的海量数据, 这对于相关内容审查工作提出了严峻挑战, 尤其是图像数据的内容审核与安全等更为困难. 但目前针对图像数据的安全分析并不成熟, 并且不法分子时常对正常运营的网站进行攻击, 将合法图像篡改为违规图像, 这严重危害网络安全. 本文针对这一实际应用需求, 设计并实现了一个面向网站图像数据的安全分析系统, 该系统主要包括以下两个模块: (1) 基于深度学习的图像内容检测引擎模块; (2) 基于事件触发技术及外挂轮询技术的图像防篡改模块. 该系统可快速审查图像数据内容是否合法并且自动监测图像数据是否被篡改.

关键词: 深度学习; 图像内容检测引擎; 图像防篡改; 安全分析系统; 图像

引用格式: 王赛赛,张磊,李健.面向网站图像数据的安全分析系统.计算机系统应用,2018,27(10):121–126. http://www.c-s-a.org.cn/1003-3254/6604.html

Security Analysis System for Web Images Data

WANG Sai-Sai[1,2], ZHANG Lei[1], LI Jian[1]

[1](Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)
[2](School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: With the advent and popularization of Internet mass media, massive data have been generated in forms of texts, images, videos, etc. This poses a serious challenge for the review of related content, especially the security review of the image data. At present, the safety analysis of image data is not mature and thus criminals often hack websites and tamper with the images on the websites, which poses significant threats to network security. Targeted at this practical application, this study designs and implements a web image data security analysis system, which is composed of two major modules: (1) illegal image detection engine module based on deep learning algorithm; (2) image tamper-proof module based on event trigger technology and plug-in polling technology. The system can quickly review whether the image data content is legitimate and automatically monitor whether the image data has been tampered with.

Key words: deep learning; illegal image detection engine module; image tamper-proof module; security analysis system; image

随着微博、知乎等网络大众媒体的出现与流行, 互联网用户规模不断扩大. 截至 2017 年 6 月, 中国网民规模已达 7.51 亿, 互联网普及率约为 54.3%. 人们分享信息的方式开始从文字、音频向图像、视频转换, 网络信息结构呈复杂化趋势. 互联网普及率的不断提高, 虽然使得人民生活更加便利, 但也带来诸多网络安全问题. 网络信息中出现很多涉及暴恐、色情和危害国家及公共安全的违规内容, 严重影响用户体验, 并且

① 收稿时间: 2018-03-14; 修改时间: 2018-04-18; 采用时间: 2018-04-27; csa 在线出版时间: 2018-09-28

对社会风气和人民生活产生恶劣影响. 违规文字较易检测, 目前已有成熟的应用软件工具. 与文字相比, 图像内容的检测更为复杂. 不法分子试图将违规内容添加在图像中以逃脱网络过滤系统的审核. 网站运营人员很早就认识到网站图像数据安全问题的重要性, 通常投入大量人力物力对图像数据进行审核管理, 这在一定程度上保证了图像数据的合法性, 但不法分子非法攻击网站、篡改图像数据时, 网站运营人员通常无法做出及时的处理.

因此, 为了能快速审查新上传至网站的图像数据是否合法并且监测网站上已存在的图像数据的合法性, 本文提出了一种面向网站图像数据的安全分析系统, 首先采用基于深度学习的图像内容检测算法对新上传的图像内容进行检测以保证数据的合法性; 其次利用基于文件监测程序的事件触发技术和外挂轮询技术实现周期性监测图像数据是否被篡改.

1 研究现状

虽然目前针对违规图像还没有成熟、高效和系统的解决方案, 但是很多研究者针对违规文本图像、色情图像和暴恐图像的检测做了大量研究, 积累了很多经验.

针对违规文本图像, 其难点在于准确地识别图像内所含文本信息. 惠普公司针对图像中的文本识别, 研发了 Tesseract -OCR 算法[1], 之后谷歌对该算法进行了改进, 算法对英文识别效果较好. 但识别包含汉字的图像时, 即使引入中文字库包 chi_sim.traineddata, 中文的识别率依然较低. 针对色情图像, Jones MJ 等研究了基于统计直方图的贝叶斯分类算法[2], 采用直方图统计肤色像素点的像素值, 进而建立肤色区域, 并利用肤色建模, 将肤色区域及人体形状性质作为参数, 采用贝叶斯分类器对其分类. 该方法简单易行, 运算效率高, 但是具有相近像素值的像素点可能会被误判为肤色区域, 因此可能产生误差, 造成分类错误. 针对暴恐图像, Paul 等开发了基于 Haar-like 特征的 Adaboost 分类器算法[3], 该算法最初应用于人脸识别领域. 利用积分图像方法和 Adaboost 分类器的特征筛选特性来提取图像特征, 再保留最有效特征, 这样可以减少运算复杂度, 同时也可提高图像检测准确率. 由于暴恐图像特征不如人脸特征明显, 所以提取的图像特征并不理想, 导致检测准确率不高. 中央民族大学研究了基于视觉语义概念的暴恐视频检测的算法[4], 其关键技术是利用视觉语义概念构建暴恐图像词频特征, 并构建视觉语义概念直方图, 然后采用 SVM 分类器进行分类检测, 根据输出结果判定是否为暴恐图像. 该算法提出根据视觉语义概念把特征分为八种类型, 若图像中包括任何一种类型的特征即判定为暴恐图像, 从而增大分类结果准确率.

以上均为针对特定单一违规图像内容的研究, 很多学者也研究了同时审查违规文本图像、色情图像和暴恐图像的检测系统. 其中, 比较有代表性的系统有: 华南理工大学学者研究了基于历史 IP 过滤的防御实验系统[5], 每次将截获到的 IP 地址更新到 IP 地址库中以增强系统的防御能力. 该方法操作简便、高效, 但缺乏实时性. 东北大学学者研究了基于语义的智能防火墙系统[6], 利用 Daubechies 小波与正则中心矩相结合的方法进行特征提取, 然后利用基于语义的特征向量匹配技术判别图像内容是否违规. 该方法具有一定的通用性, 但不具备信息反馈功能, 缺乏自适应性. 在基于语义特征的基础上, 将动态知识库和规则库一起作为特征向量匹配的参考因素, 西南交通大学学者提出了基于知识的智能网络安全监测系统[7], 该方法具有信息反馈和知识导航功能, 但采用小波分析和特征不变量的方法进行特征提取, 得到的特征维度较低, 不能很好地表达特征.

以上三种系统均采用传统机器学习方法提取特征, 而传统机器学习模型均为浅层结构, 并且大多依赖先验知识, 擅长分析维度较低的数据. 面对高维的图像数据, 运用具有深层结构的深度学习模型, 自动学习提取高维特征向量, 更有效率, 同时表达能力也更强. 因此, 本文采用基于深度学习的图像内容检测引擎来完成新图像数据的审核工作.

网页防篡改技术[8]是保障网页内容安全的一种技术, 目前应用较多的技术有: 1) 外挂轮询技术: 基于文件检测程序, 以轮询方式将被监控文件与相对应的文件比较, 然后判定文件是否被篡改. 2) 事件触发技术: 利用操作系统的文件系统或驱动程序接口, 在网页文件被修改时进行合法性检查. 外挂轮询技术实现简单, 不需要有管理员权限且具有自我防护能力, 但轮询时间周期设定过长, 则缺乏时效性; 若轮询周期设定过短, 频繁调度则降低服务器性能. 事件触发技术具有很强的自我防护能力, 能够防范连续篡改攻击, 能够实时检测事件是否发生, 服务器负载低, 因而对网站访问性能

影响较小, 但需要对所监测网站的文件系统有一定的访问控制权限. 本系统尝试将上述两种网页防篡改技术应用到网站图像数据防篡改上.

2 系统总体架构

系统采用功能模块化设计结构, 降低系统复杂度, 使系统设计、调试和维护等操作简单化, 具有更多灵活性, 便于后续根据实际需求进行调整或扩展. 本系统中, 图像内容检测引擎针对特定违规图像内容分别运用不同的算法进行检测, 力求达到最高的运算效率和识别精度, 并且加入图像防篡改模块保证网站图像数据的合法性. 系统总体架构如图 1 所示.

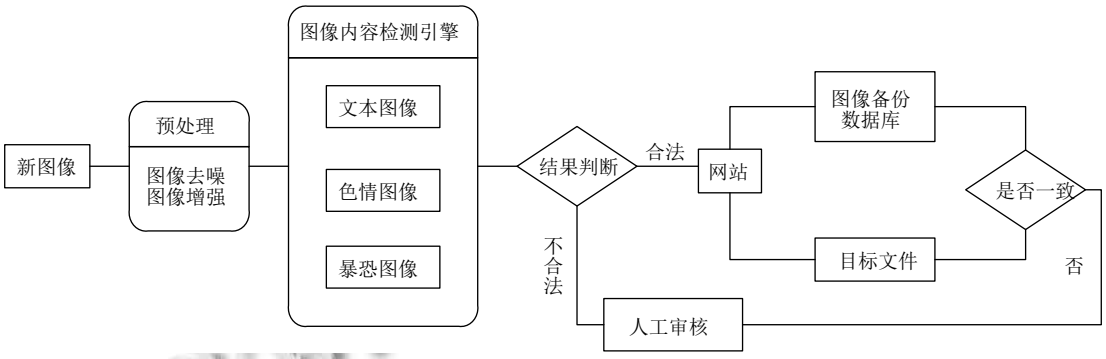

图 1 面向网站图像数据的安全分析系统总体架构

图像预处理模块: 对新图像进行图像去噪、图像边缘增强等操作, 去除图像中无关信息、增强边缘信息以便增加图像特征信息的可检测性.

图像内容检测引擎模块: 运用深度学习算法对新图像数据进行图像内容检测, 审查是否存在违规内容. 若图像数据合法, 则提取其 MD5 编码到图像备份数据库; 否则将图像数据提交给人工审核. 该模块的功能是保证新上传至网站的图像数据的合法性.

图像防篡改模块: 利用事件触发技术和外挂轮询技术对网站中所有图像数据进行监测, 若检测到图像数据异常, 则移交人工审核. 该模块的功能是监测网站中已存在的图像数据合法性.

图像内容检测引擎模块和图像防篡改模块实现了本系统的核心功能, 下面将对两个模块中所用到的关键技术做具体分析和介绍.

3 关键技术分析

图像内容检测引擎包括违规文本图像检测、色情图像检测及暴恐图像检测三个子模块. 经检测后的图像数据, 若结果判断为合法, 则将其上传至正常运营的网站后, 采用网站图像数据防篡改策略对该图像数据进行监测, 检测其是否处于安全状态.

3.1 违规文本图像检测

文本图像检测类似于目标检测, 先检测到目标区域, 然后进行特征提取、特征匹配及结果输出. 但不同之处在于特征匹配. 对于文本检测, 同一个文本线的多个字符组成一个序列, 线上不同字符可以互相利用上下文信息. 本系统采用卷积神经网络 (CNN) 对违规文本图像进行特征提取, 然后将同一文本线的字符特征输入到长短时记忆网络 (LSTM) 中进行分析与综合, 最后将分类得到的目标区域合成文本线并将最终文字作为结果输出. 这种把 CNN 和 LSTM[9]无缝结合的方法提高了检测精度和效率.

违规文本内容检测模块可以根据上述思路自行编码实现, 也可以利用现有开源代码及开放接口进行定制开发, 本系统采用第二种方式, 具体流程如图 2. 首先, 待检测图像利用百度 AI[10]开放的 basicGeneral 接口识别出图像中所含文字. 然后, 采用基于 TextRank 算法的 jieba.analyse.textrank 接口[11]对其提取关键词信息. 最后, 将提取的关键词同敏感词汇库做匹配. 若匹配成功, 则该图像属于非法图像, 返回人工审核; 否则继续进行色情图像检测. 其中, 关键词匹配采用模板匹配[12]与正则表达式相结合的方法.

3.2 色情图像检测

该模块采用何凯明等人提出的深度残差网络[13]来解决图像二分类问题, 将训练图像数据输入到 ResNet 网络中进行训练, 最终根据模型将待检测图像数据分为合法图像和非法图像. 考虑到色情图像数据集性质

问题, 本系统采用雅虎公司开源项目 open_nsfw[14]中已训练完成的 resnet_50_1by2_nsfw.caffemodel 模型. 该模型采用 ImageNet+NSFW(Not Suitable For Work) 数据集, 首先在 ImageNet 1000 类数据集上进行预训练, 然后根据 NSFW 数据集对权重进行调整. 该方法采用的是 pynetbuilder 工具生成的 resnet50_1by2 作为预训练网络. 虽然更深层的网络或者采用更多滤波器的网络可能会提高准确率, 但综合考虑准确率、训练时间及参数问题, 最终选择 resnet50_1by2 网络作为系统预训练网络, 最后一层为 softmax 层, 可以输出图像在合法、非法两类中各自的概率值. 模型以 SFW 值 (合法图像类) 为判定依据: 当 SFW 值的概率大于 0.8 时, 判定该图像是合法图像; SFW 值低于 0.2 时, 判定为违法图像; SFW 值处于 0.2–0.8 之间时, 图像合法性随着 SFW 值减小而降低, 判为不确定图像.

系统将 SFW 值大于 0.8 的合法图像送入暴恐图像检测模块继续检测, 对于 SFW 值小于 0.8 的违法图像及不确定图像, 则移交给人工审核.

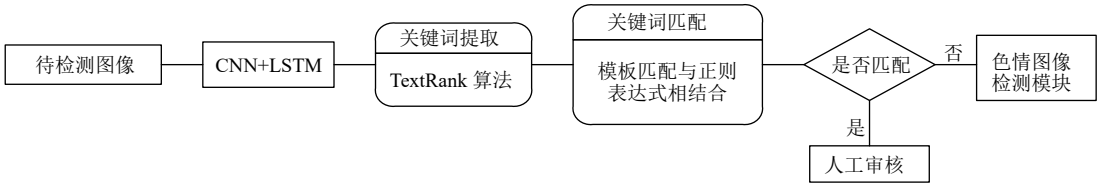

图 2 违规文本内容检测模块流程

3.3 暴恐图像检测

根据视觉语义概念, 可以将暴恐图像数据集分为以下九类: 正常图像、爆炸火灾图像、暴乱图像、血腥图像、军事武器图像、杀人图像、尸体图像、暴恐人物图像以及警察部队图像. 鉴于传统机器学习方法擅长分析维度较低的数据, 而深度学习算法擅长分析高维度的数据, 本系统采用卷积神经网络对数据集进行特征提取和模型训练, 判别待检测图像分别属于九类图像的各自概率值, 最后综合每类概率得出图像总概率值, 根据概率值判定待检测图像是否为暴恐图像. 鉴于数据集收集比较耗费人力和时间, 最终本系统采用百度 AI[11]开放的 antiTerror 接口对暴恐图像进行图像内容检测, 结果输出为各项概率值. 其中, 合法图像数据的综合概率值均在 0.9 以上.

3.4 网站图像数据防篡改策略

图像防篡改模块采用事件触发技术[15]或外挂轮询技术予以实现, 具体技术原理如图 3 所示. 当合法图像上传到网站时, 自动提取图像的 MD5 编码[16], 并且存储于图像备份数据库. 随后可以根据是否拥有对网站文件系统具有访问控制权限而采取不同方式:

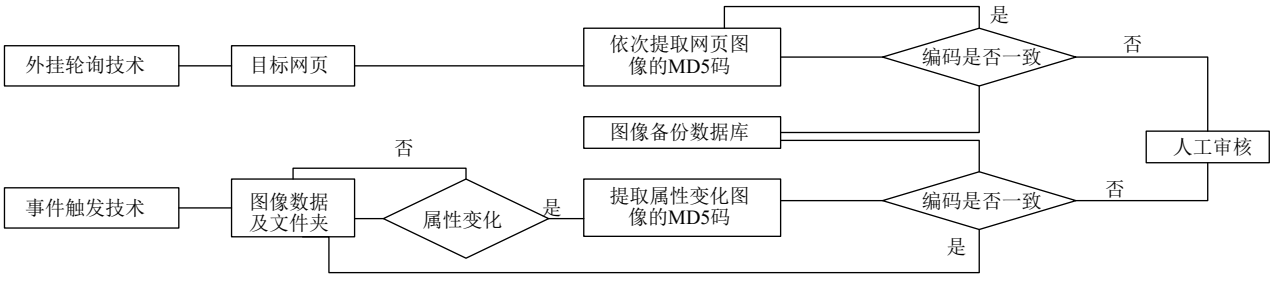

图 3 图像防篡改模块流程图

1) 若有文件系统访问控制权限, 可以利用事件触发方式对图像数据及文件夹进行实时监控, 若图像文件属性发生变化, 则将提取该图像数据的 MD5 编码, 然后与图像备份数据库中对应图像的 MD5 编码比对, 检测已上传的图像数据是否被恶意篡改. 若 MD5 编码不同, 则表示图像数据异常, 应将图像信息反馈给人工审核. 本策略同样适用于非图像类文件的防篡改.

2) 若无文件系统访问控制权限, 则采用外挂轮询

技术周期性地从外部逐个访问目标网页中的图像数据, 提取图像数据的 MD5 编码与真实的图像数据的 MD5 编码相比较, 若发现异常, 则移交人工审核.

4 实验分析

本次实验程序主要采用 Python 语言编写, 运行在带有 16G 显存容量的 Ubuntu 16.04 系统的服务器. 新图像进入图像内容检测引擎模块后, 首先进行以下预处理操作:

1) 利用小波域高斯混合模型[17]对图像去噪处理.

2) 将图像统一转换为 RGB 图像.

3) 对图像进行边缘增强处理.

通过完成图像去噪、图像增强等操作以达到去除图像中无关信息, 并增强边缘信息、突出图像特征的目的. 以下为违规文本图像和色情图像检测的实验结果.

4.1 违规文本图像内容检测

本次实验所用图像均来自合法公开的图像发布平台, 所有图像总共包括汉字 2066 个, 英文字母 2522 个.

表 1 Tesseract-OCR 和 basicGeneral 方法识别正确率 (%)

文字类型	Tesseract-OCR	basicGeneral
英文	97.2	97.8
汉字	66.4	82.3

从实验中可以得到结论: 两种方法对英文识别率均比较高, 其中, basicGeneral 方法效果略好于 Tesseract-OCR 方法. 对于汉字的识别, 整体准确率不如英文识别率高, 但 basicGeneral 方法的准确率明显高于 Tesseract-OCR 方法, 本系统中采用 basisGeneral 方法.

4.2 色情图像内容检测

实验所用图像来自于公开的图像数据集 ImageNet[18]. 图 4(b) 是用利用直方图提取的肤色特征, 根据贝叶斯分类算法, 判定为非法图像, 而该图像显然为合法图像, 故分类错误. 而在本系统中, 根据 NSFW 算法可计算得到该图像的 SFW 值为 0.99806279. 该值大于 0.8, 根据 3.2 部分的判别规则, 算法判别为合法图像, 分类正确. 因此, NSFW 算法更适合色情图像内容检测.

本次实验从 ImageNet 数据集中随机选取 1165 张图像数据作为色情图像检测和暴恐图像检测的测试图像数据集. 其中, NSFW 算法将 19 张合法图像误判为非合法图像, 误判率为 1.63%; 暴恐图像检测算法将 26 张合法图像误判为非合法图像, 误判率为 2.23%. 综上, 误判率均在系统误差可接受范围.

(a) 原图像

(b) 运用直方图统计的肤色特征图

图 4 色情图像检测结果

5 结束语

本文提出的网站图像数据安全分析系统, 不仅可以对图像内容进行检测, 及时检测出违规文本图像、色情图像和暴恐图像等, 而且可以对图像数据进行周期性或者事件触发式的监测以防止非法篡改, 因此主要包括图像内容检测引擎模块和图像防篡改模块. 图像内容检测引擎模块针对三类违规图像内容分别采用开源的 basicGeneral 接口、NSFW 算法以及 antiTerror 接口进行定制开发, 综合了各个算法对不同图像的识别优势, 得到了较好的集成效果, 违法图像识别准确率较高, 可以检出大部分非法图像, 同时辅以对不确定图像的人工审核, 兼顾了检测的时效性和准确性, 使该系统具备了较高的实用性. 图像防篡改模块可以监测图像数据是否安全, 即使发生黑客入侵网站并篡改图像内容, 也可以及时发现并妥善处理.

参考文献

1 Smith R. An overview of the tesseract OCR engine. Proceedings of the 9th International Conference on Document Analysis and Recognition. Parana, Brazil. 2007. 629–633.

2 Jones MJ, Rehg JM. Statistical color models with application to skin detection. International Journal of Computer Vision, 2002, 46(1): 81–96.

3 Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. Proceedings of 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Kauai, HI, USA. 2001. 511–518.

4 宋伟, 杨培, 于京, 等. 基于视觉语义概念的暴恐视频检测. 信息网络安全, 2016, (9): 12–17. [doi: 10.3969/j.issn.1671-1122.2016.09.003]

5 简校荣. 基于历史 IP 过滤的防御实验系统研究与实现[硕士学位论文]. 广州: 华南理工大学, 2013.

6 许强, 江早, 赵宏. 基于图像内容过滤的智能防火墙系统研究与实现. 计算机研究与发展, 2000, 37(4): 458–464.
7 罗小宾, 魏万迎, 夏晓东. 基于图像内容过滤的智能监测系统的研究. 微计算机信息, 2007, 23(27): 71–72, 32. [doi: 10.3969/j.issn.1008-0570.2007.27.028]
8 Shao H, Yu TS, Xu MJ, *et al*. Image region duplication detection based on circular window expansion and phase correlation. Forensic Science International, 2012, 222(1-3): 71–82. [doi: 10.1016/j.forsciint.2012.05.002]
9 Tian Z, Huang WL, He T, *et al*. Detecting text in natural image with connectionist text proposal network. In: Leibe B, Matas J, Sebe N, et al, eds. Computer Vision – ECCV 2016. Cham: Springer. 2016. 56–72.
10 百度 AI. https://cloud.baidu.com/product/imagecensoring.
11 Mihalcea R, Tarau P. TextRank: Bringing order into texts. Proceedings of EMNLP. Barcelona, Spain. 2004. 404–411.
12 Li ZH, Liu CY, Cui JG, *et al*. Improved rotation invariant template matching method using relative orientation codes. Proceedings of 30th Chinese Control Conference. Yantai, China. 2011. 3119–3123.
13 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 770–778.
14 雅虎. Open nsfw. https://github.com/yahoo/open_nsfw.
15 冶忠林, 王相龙. 网页防篡改和自动恢复系统. 计算机系统应用, 2012, 21(2): 225–228. [doi: 10.3969/j.issn.1003-3254.2012.02.053]
16 张学旺, 唐贤伦. MD5 算法及其在文件系统完整性保护中的应用. 计算机应用, 2003, 23(Z2): 430–432.
17 胡晓东, 彭鑫, 姚岚. 小波域高斯混合模型与中值滤波的混合图像去噪研究. 光子学报, 2007, 36(12): 2381–2385.
18 Deng J, Dong W, Socher R, *et al*. ImageNet: A large-scale hierarchical image database. Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA. 2009. 248–255.