移动应用图片文字识别系统

刘佳成、杨添宝

2019 年 9 月 30 日

1 项目背景

数据显示，截至 2017 年国内安卓移动应用市场中充斥超过 400 万款移动该应用，国内移动终端用户数量超过 10 亿，移动互联网俨然成为网络信息传播的一个庞大而有效的渠道，为社会生活提供了极大的便利，同时，为违规/违法信息传播提供了便利。

对于移动互联网中传播的违规/违法信息，国家监管部门需要进行管控，找到并对相关信息查处。此类违规信息在移动互联网中存在形式有多种，可能是文字、图片、链接、音频、视频等。其中文字违规信息识别市场中已有较为成熟的语义识别工具，能够准确的将违规/违法信息过滤与屏蔽，对于违法链接也可通过建立黑名单的形式一定程度上解决，但对于图片、音视频类型违规信息目前市面上并没有成熟、全面的解决方案。

为逃避监管机构的监管，违法信息发布者通常会将违规文字信息制作成图片的形式在应用中发布，增加了监管机构的识别难度。

2 项目介绍

移动应用图片文字识别系统，通过获取移动应用中的图片资源文件，逐个扫描图片文件，识别并存储图片中出现的文字内容，进一步可通过语义识别工具判别文字内容是否为不良信息，从而帮助监管机构筛选出移动互联网中传播的违规/违法图片信息。

3 业务场景

某地网络信息监管部门获取了一款疑似在互联网中散播有害社会舆论的违规图片信息的应用，为采集违规信息证据，将应用提交到了移动应用图片文字识别系统中。系统通过模拟运行该款应用，获取了应用中的内置图片资源及运行时产生的图片资源，并识别、打印出了图片资源文中的文字信息，再通过语义识别系统对文字信息进行了排查，最终找出了该款应用中可疑的违规内容。

监管部门整理了该款应用中散播的有害图片信息，包括信息存放位置、内容、存储格式等，收集并整理到了应用违规检测报告中，成功对该款应用运营商与发布了该应用的渠道商下发了整改通知。

4 项目内容

1. 该项目将使用 B/S 或 C/S 架构设计完成一个应用违规信息检测系统，该系统能够对应用中的图片和文字信息进行提取，使用 OCR 技术对图片中的文字进行识别，通过一定的关键字及 AI 语义分析方法识别文字中是否包含暴力、色情、赌博、血腥、宗教、党政等敏感信息并标记带有违规信息的内容。

2. 该系统将会支持包括 jpg、png、icon、jpeg、gif、bmp 在内的多种图片格式，将图片中包含的所有文字信息（主要是各种字体的机打文字）识别并在系统内输出展示，同时检测图片中包含的违规信息，尽量做到具有较高的识别率。

3. 该系统将会支持应用内的文字提取，包括资源文件中文字和应用运行过程中显示的文字，同时支持模拟运行应用并记录应用运行过程中从网络获取的图片以及运行过程中的截图，针对图片和文字进行检测违规信息。

4. 该系统将会具有良好的交互能力和使用体验，支持设定违规信息关键字、支持对单独上传的图片进行检测、支持批量导入应用和图片进行检测。

5. 该系统能够生成一定格式的检测报告供以后浏览，并且能够对不同应用的检测结果存储保存、对结果进行快速检索。

5 项目计划

十月上旬 学习 Android APK 应用的结构，了解如何提取其中的资源文件等信息、确定项目的架构。

十月中下旬 完成对纯文字信息中敏感信息的识别，熟悉 B/S 架构或 C/S 架构程序的设计方法。

十一月 完成对图片中的文字的提取、掌握模拟运行应用并提取其在网络中获取的图片信息的方法。

十二月 将各个部分整合并完成整个系统，优化交互能力和识别效果等。