DOI:10.16515/j.cnki.32-1722/n.2019.04.005

一种基于众包的数据标注系统

朱子健[1],蔡树林[2,3],张墨轩[1],施佳佳[1],赵海峰[3*]

(1.金陵科技学院国际教育学院,江苏 南京 211169;2.浙江大学软件学院,浙江 宁波 315000;
3.金陵科技学院软件工程学院,江苏 南京 211169)

摘 要:构建了一个基于众包的数据标注系统,用于大规模数据的人工实时并行标记与验证。主要系统角色为管理员和标注用户。管理员的主要功能为用户管理和图片集管理,标注用户的主要功能为个人信息管理与数据标注管理。系统主要采用SSM框架来实现业务逻辑,通过使用MyBatis对JDBC进行封装,实现对MySQL数据库操作。系统前端实现采用JSP动态页面与Bootstrap框架。通过网络爬虫抓取互联网上的各类图片构建测试数据集,对系统进行了测试,验证了系统的可行性。

关键词:众包;数据标注;SSM;Web开发

中图分类号:TP391 **文献标识码**:A **文章编号**:1672-755X(2019)04-0020-05

A Crowdsourcing-based Data Annotation System

ZHU Zi-jian[1], CAI Shu-lin[1,2], ZHANG Mo-xuan[1], SHI Jia-jia[1], ZHAO Hai-feng[1*]

(1. Jinling Institute of Technology, Nanjing 211169, China; 2. Zhejiang University, Ningbo 315000, China)

Abstract: In this paper, we propose a crowdsourcing-based data annotation system, which can be used for real-time manual labelling and verification of large-scale data. The main roles in the system are administrators and annotated users. The main functions of the administrator are user management and image management. The main functions of the annotation users are personal information management and data annotation management. We adopt SSM framework to implement the core logic of the whole system. We implement the operation of the MySQL database by encapsulating JDBC with MyBatis. The front-end of the system is implemented using JSP and Boostrap. The test image dataset is captured by crawlers. The tests are conducted to verify the effectiveness of the system.

Key words: crowdsourcing; data annotation; SSM; web development

2012年深度学习在ImageNet挑战赛上取得重大突破,使其受到了越来越多的关注,在图像、语音、文本等领域中广泛应用,并在人脸识别、自动驾驶、智能助理等方面成功落地,掀起了人工智能的新一轮高潮。深度学习取得突破的原因之一是其能够通过自动从数据中学习具有高度鉴别能力的特征表示。相比于传统机器学习需要手工设计特征表示方法,深度学习的特征表示学习能够更好地捕捉到数据的内在规律,因此具有更好的泛化性能。另一方面,深度学习将传统的神经网络的层数不断加深,从AlexNet[1]的8层、VGGNet[2]的19层,到ResNet[3]的152层,甚至到目前的成千上万层,不断加深的网络层数能够学习

收稿日期:2019-09-26

基金项目:江苏省高校自然科学研究重大项目(16KJA520003);金陵科技学院孵化基金(jit-fhxm-201808);金陵科技学院高层次人才引进基金(jit-b-201717);江苏省高等学校大学生创新创业训练计划(201813573015Z)

作者简介:朱子健(1997—),男,江苏连云港人,主要从事软件工程方面的研究。

通信作者:赵海峰(1984—),男,河南三门峡人,高级工程师,博士,主要从事计算机视觉方面的研究。

到更加复杂的模型,也就具有更强的泛化能力。然而,无论是表示学习,还是更深的网络结构带来的更复杂的模型,都需要大量的数据作为输入。否则,数据的过拟合问题会导致深度学习无法学习出好的特征表示,也无法得到复杂度较高的模型,从而影响模型的泛化性能。

随着行业应用系统的规模迅速扩大,行业应用所产生的数据呈爆炸性增长,动辄达到数百 TB 甚至数十至数百 PB 规模。百度目前的总数据量已超过 1 000 PB,每天需要处理的网页数据达到 10～100 PB;淘宝累计的交易数据量高达 100 PB;Twitter 每天发布超过 2 亿条消息;新浪微博每天发帖量达到 8 000 万条;中国移动在中国一个省的电话通联记录数据每月可达 0.5～1 PB;一个省会城市公安局道路车辆监控数据 3 年可达 200 亿条、总量 120 TB。据世界权威 IT 信息咨询分析公司 IDC 研究报告预测:全世界数据量到 2020 年将为 35 ZB(1 ZB=1 024 EB),年均增长 40%。

当前,海量数据大都以无标签的形式存在。目前为止,深度学习中最有效的方法仍然属于监督学习的范畴,需要大量的标注好的数据作为训练样本,以便从中学习出有效的模型,从而对未知的样本进行推理预测。数据标注,即对数据按照学习任务的不同加上相应的标签,成为深度学习过程中一个不可或缺的重要环节。由于数据量太大,直接人工数据标注一方面需要花费大量的人力和时间成本,另一方面由于长时间工作,疲劳现象不可避免,数据标注的准确度会大大下降,从而导致机器在错误的数据标签中无法学习到有效的信息,最终影响深度学习模型的泛化能力。

本文针对当前深度学习中大规模无标签数据亟待高效标注的问题,提出了以众包形式来进行大规模数据标注的方法,设计开发了一个基于众包的深度学习数据标注系统。该系统针对不同的任务,能够为用户提供一套用于数据标注的工具,包括针对图像识别问题的类别标签标注工具,针对目标检测问题的矩形、多边形标注工具,以及针对图像分割问题的目标轮廓标注工具。同时,系统还提供了像素值提取等辅助工具。通过该系统,多名标注人员可以同时对数据进行标注。

1　数据标注的国内外研究现状

自机器学习发展以来,数据标注都是机器学习的必要前提。在数据规模不大的时候,一般都是自行标注。例如,在做细胞分类的时候,需要人工事先将不同的细胞类别标注出来,以便机器学习到不同类别的细胞。小规模的数据,一般由单人标注或者几个人标注即可完成,标注人员往往由实验人员临时担任。

随着机器学习的不断深入和广泛应用,数据规模不断增加,要解决的问题也不断增多,越来越多的研究人员从事这一领域。为了比较不同的方法,一些研究小组将其论文中使用的数据发布出来,共享给整个领域使用,形成了专门的共享数据集。比较有代表性的是美国加州大学欧文分校的 UCI Machine Learning Repository[4],其收集了 429 个机器学习领域的数据集,包括著名的 Iris 数据集,涵盖了机器学习的多个子领域。然而,这些数据规模仍然比较小,适合小规模的机器学习。

2003 年,Li F. F. 等收集了被称为 Caltech101 的图像数据集[5],其包含了 101 个物体目标类别和 1 个背景类别,每类从 40 到 800 张图片不等,大部分的类别含有 50 张左右的图片。Caltech101 数据集一共大约有 1 万张图片,覆盖了平日常见的物体目标。与之前的数据集不同的是,Caltech101 的数据集有了一定的规模(数量已经上万)。同时,研究人员对数据进行了精细标注,标注了每个物体的精确轮廓,并且数据集上提供了相应的工具来标注和显示。自发布后,Caltech101 成为图像分类和语义分割领域最有效的数据集,大量的算法在 Caltech101 上得到验证,大大促进了计算机视觉技术的发展。到 2006 年,又推出了 Caltech256 数据集[6],其包括了大约 3 万张图片,256 类。标注方法与 Caltech101 类似,也是精确的轮廓和类别标注。

随着数据集规模和数量的不断增加,已经无法单独采用直接人工标注方法来进行数据标注。Torralba 等提出了 LabelMe 数据标注工具[7],其采用在线的方式对图像的轮廓、类别、目标名称等进行标注。该工具包括了二维轮廓、精细类别以及标注编辑等一系列工具。从 2003 年起,LabelMe 已经先后标注了超过 1 000 万张图片。

在数据标注工具的催生下,欧洲的 PASCAL 项目推动了名为 Visual Object Challenge(VOC)的比

赛[8]，2005—2012年先后举办了8届，取得了巨大成功。随后，大规模可视识别挑战赛(Large Scale Visual Recognition Challenge，ILSVRC)[9]举办，此比赛以ImageNet作为基础数据。ImageNet包含百万量级的图片，采用了亚马逊的Amazon Mechanical Turk(AMT)，以众包的形式为ImageNet提供大量的高效标注，雇佣了超过2.5万标注人员，在2年内标注了超过1 000万张图片以及1万量级的语义信息。通过采用众包方法，大大提高了数据标注的效率，也保证了数据标注的准确度，为深度学习技术的发展奠定了数据基础。

国内早在2006年，湖北鄂州莲花山研究院就已经开始了专人进行数据标注的工作。数据标注师通过标注，不仅提供类别等信息，更为图像提供语义信息，建立一个基于语义的网络结构，作为早期的图像语义网。当前，国内著名的标注公司如数据堂等，为特别需求提供定制化的数据标注服务。然而，基于众包的数据标注还处于较为初级的阶段。

从国内外发展趋势来看，海量规模的数据必须需要相应高效的数据标注平台，而基于众包等技术的平台是解决数据标注的有效方案之一。

2 系统总体架构

众包数据标注系统采用典型的浏览器-服务器即B/S结构来实现。B/S结构的好处之一是其客户端不依赖于特定的操作系统平台，而只依赖于浏览器。由于在不同的操作系统平台上都有浏览器，因此，系统也就实现了跨平台功能。系统的总体架构如图1所示。从图中可以看到，系统由客户端和服务器端两部分组成。客户端主要负责数据标注的交互部分，包括了系统管理员的界面和标注用户界面；服务器端则是负责对数据的实际处理和存储。

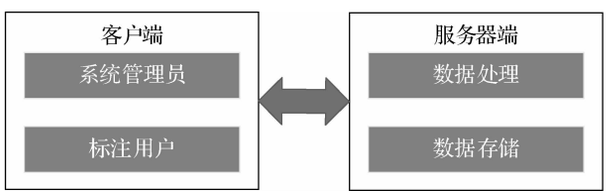


图1 系统总体架构

系统管理员主要负责对系统管理操作，其对数据库中各类对象的操作如表1所示。标注用户主要对用户本身进行操作，其对数据库中的各类对象的操作如表2所示。

表1 管理员端的功能设计

操作对象	具体功能
系统对象	进行登录信息的验证
	退出
标注用户	展示所有的普通用户的信息
	可以进行条件查询用户信息
	可以实现用户信息的修改、删除、创建
图片集合	展示所有的图片集合信息，同样进行分页处理
	可以进行条件查询集合信息
	修改集合信息
	可以实现图片信息的上传，并包括标记类型的设置
	可以实现图片集合的下载，并包括在线压缩
图片	将集合中图片的信息提取到数据库中

表2 标注用户端的功能设计

操作对象	具体功能
系统对象	登录系统
	退出系统
	注册功能
用户本身	查看个人信息
	修改个人信息
图片集合	实现查看集合
图片	进行图片标记功能
	对当前文件和标记进行下载
	进行图片大小的缩放

3 系统的具体实现

本节主要详细说明系统中的标注逻辑以及图片集合的格式管理的具体实现，这两个部分是系统的关键。其中，图片标注主要针对的是图像分类、检测和分割问题。主要任务包括对图片添加目标轮廓的矩形框、标注出轮廓的完整曲线以及对图片添加类别标签等。此外，由于数据规模较大，需要对于图片进行管理，建立结构化的数据集，以便提高系统的整体性能。

1)图片标注实现。在众包框下，一张图片需要多名用户对其进行标注。由于每位用户的知识背景不

同,对图片的标注可能存在偏差。为了提高标注的准确率,本文从业务逻辑入手,实现“一张图片多次标注检查,问题上抛”的操作。对于每张图片,都经过两位不同的用户进程标注,一旦发现其中的标注信息有误,发现者可以对该处的标注进行标记,则此时图片的状态标记为 Error,该图片会进入另外两位用户手中,重新进行标注,如果还存有问题,则将该图片发送给管理员,执行最终标注。

2)图片集合的处理。针对流读取和字符解析过程中格式的一致性问题,系统对图片集合的制作格式,采取如下方法:①根据文件的目录结构,对于上传到服务器的图片集合 zip 进行解压处理。②将关键字为.txt 的文件内的具体内容进行数据读取。③将图片文件夹中的图片信息,进行遍历存储。在此基础上,对图片进行统一存储。

3)标注信息的格式。对于标注信息内容,本文存储对角像素位置,以及具体的标注信息内容。其在数据库中存储的结构形式为:

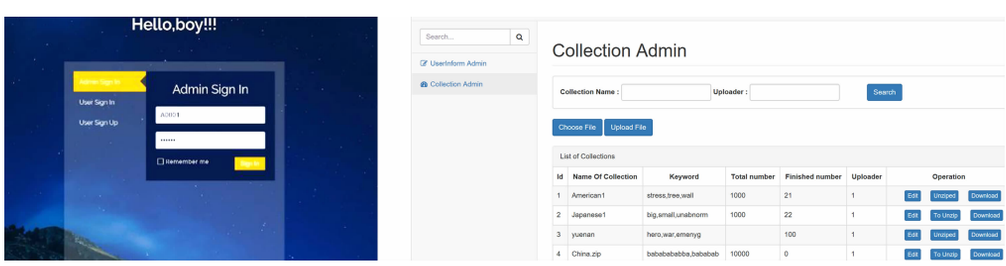
[{"x":0.33258843680203265,"y":0.20846800258564965,"ex":0.5878721274315304,"ey":0.546218487394958,"name":"urnrfni"},{"x":0.5212105108873293,"y":0.795087265675501,"ex":0.776494201516827,"ey":0.9647705235940531,"name":"nirnfir"}]

对于整个系统来说,由于图片为二进制数据,因此以文件形式进行存储。而对于图片标注的内容,则采用数据库的形式进行存储,这样大大提升了数据的存取效率。

4　实验结果及分析

4.1　系统部署

系统采用 MySQL 数据库,使用 Tomcat 进行部署。具体操作时,将系统 WAR 包拷贝到 Tomcat 相应目录下,运行即可部署完毕。绑定外网域名,即实现处处可访问。图 2 是用户登录注册和数据收集界面。主要功能包括管理员登录、用户登录、用户注册。其中管理员登录信息只能在数据库中进行修改。用户登录注册信息主要由用户自己进行注册。用户可以自行修改个人信息。管理员具有修改用户信息的权限。



(a)用户登录界面　　(b)数据收集界面

图 2　用户登录和数据收集界面

4.2　管理员端的使用与测试

通过测试,管理员在依据上述功能实现操作后,所有的功能均能实现。

4.3　标注用户端的使用与测试

通过网络爬虫在互联网上爬取多种图像,制作图像数据集。标注用户对图片进行多种功能标注,能够满足图片的基本需求。图 3 为图像标注界面。

(a)图像标注界面　　(b)结果界面

图 3　图像标注和结果界面

5 结 语

本文设计并实现了一种基于众包的数据标注系统，通过采用客户端与服务器端的实现形式，使得系统能够跨平台使用。在实际使用中，标注任务发布后，还可对每位用户所做出的贡献进行统计和计算，建立标注激励模型。此外，本文实现的标注系统仅限于标注图片，在后续工作中，将对更多数据类别标注进行研究。

参考文献：

[1] Krizhevsky A，Sutskever I，Hinton G. ImageNet classification with deep convolutional neural networks[C]. Lake Tahoe：Proceedings of the 25th NIPS，2012：1097－1105

[2] Simonyan K，Zisserman A. Very deep convolutional networks for large-scale image recognition[C]. San Diego：Prceedings of ICLR，2015

[3] He K，Zhang X，Ren S，et al. Deep residual learning for image recognition[C]. Las Vegas：Proceedings of the IEEE Conference on CVPR，2016：770－778

[4] Dua D，Graff C. UCI machine learning repository[EB/OL]. (2019－03－18)[2019－08－31]. http://archive. ics. uci. edu/ml

[5] Li F F，Fergus R，Perona P. Learning generative visual models from few training examples：an incremental Bayesian approach tested on 101 object categories[J]. CVIU，2005，106(1)：59－70

[6] Griffin G，Holub A，Perona P. Caltech-256 object category dataset[M]. California：Caltech Technical Report，2017

[7] Russel B，Torralba A，Murphy M，et al. LabelMe[EB/OL]. (2018－05－07)[2019－08－31]. https://github. com/CSAILVision/LabelMeAnnotationTool

[8] Everingham M，Van Gool L，Williams C，et al. The PASCAL visual object classes(VOC) challenge[J]. International Journal of Computer Vision，2010，88(2)：303－338

[9] Russakovsky O，Deng J，Su H，et al. ImageNet large scale visual recognition challenge[J]. IJCV，2015，115(3)：211－252

（责任编辑：湛 江）