基于 Flask 与爬虫技术的可视化深度学习数据标注系统

赵北庚

（中国刑事警察学院网络犯罪侦查系，辽宁沈阳，110035）

基金项目：中央高校基本科研业务费专项资金资助（项目编号：D2019023）。

摘要：近年来，基于深度学习的人工智能技术成为计算机科学领域的研究热点。训练和评估深度学习模型需要高质量的大规模数据集。数据集的构建过程很大程度上依赖于人工手动标注，数据标注师所标注的标签质量决定了能否训练出表现优良的深度学习模型。本文提出一种基于Flask框架和爬虫技术的可视化深度学习数据集标注系统，该系统呈现给数据标注师友好易用的标注界面，并通过基于网络爬虫技术的方法在标注时给出提示，可有效提高数据标注师对图文数据的标注质量。

关键词：数据标注；深度学习；数据集构建；网站开发

DOI:10.16589/j.cnki.cn11-3571/tn.2020.20.015

0 引言

基于深度学习的人工智能技术越来越成为计算机科学领域的研究热点。相关的研究已经被广泛应用于商业，医疗，军工，民生等各个领域。训练深度学习模型需要大规模的数据集，而构建数据集很大程度上依赖于人工标注。近年来，随着深度学习研究的火热以及对大规模数据集需求的激增，催生了一种新的生态产业—数据标注，以及相关的新职业—数据标注师。然而，由于数据标注师的水平参差不齐，所标注的数据标签质量也有高低之分。研究在标注过程中提升数据标注师标注质量的方法，对提升深度学习模型的训练和评估具有积极意义。

随着软件工程技术的发展，数据标注平台的架构和实现方式也在发生变化。早期的数据标注平台通过单独的客户端软件搜集标注师的标签，并通过服务器端实现同步。随着网站开发技术的进步和网站开发框架的普及，越来越多的数据标注平台通过网站的形式呈现给数据标注师，使得标注师可以使用浏览器进行标注。然而，相比由新型网站框架所开发的标注平台，由传统网站框架开发的标注平台网站具有界面不友好、不易使用、开发效率低、难以扩展和维护等局限性。

基于此，本文提出一种基于 Flask 框架和网络爬虫技术的深度学习数据标注平台。Flask 框架具有敏捷，高效，易于扩展和维护的特性。另外，Flask 框架以 Python 语言为基础，可以更有效地和众多以 Python 语言为基础的第三方深度学习程序库协同工作。同时，本文所提出的数据标注平台通过网络爬虫技术预先爬取有价值的标签信息，以辅助提示信息的方式呈现给标注师，可以有效地提升数据标注师的标注速度以及标签质量。

1 需求分析与设计

本文所提出的标注平台力求从以下两方面提高数据的标注质量：首先，在标注过程中，应提供给数据标注师有价值的辅助提示信息；其次，标注平台须以友好易用的形式呈现给用户，且平台本身应易于维护和扩展。为实现以上两方面目标，细化了数据标注平台的具体需求分析。重点的需求分析如表 1 所示。

表1　重点需求分析

需求	描述
高效易维护	平台应采用高效易维护的框架开发与实现，降低开发与维护成本。
适合深度学习	平台应采用适合深度学习的语言与框架进行开发，方便后续与其他第三方深度学习程序库对接使用。
辅助提示	提供给数据标注师有价值的辅助提示标注信息，所提示的信息须能够提升标注标签的质量。
标签存储	标签数据应以便于移植，便于扩展的标准化格式进行存储。
界面友好	数据标注界面应以简洁清晰的方式呈现给用户，且界面中的控件图标和文字说明应易于用户理解。

基于表 1 所述的需求分析，提出如下所述的设计方案。采用基于 Flask 的网站开发框架实现标注平台的基础功能。Flask 框架不仅高效易用，且开发过程采用 Python 语言，可以方便有效地与第三方 Python 深度学习程序库联合使用。在此基础上，采用基于 Python 的网络爬虫模块，解析待标注的图像数据并调用第三方网络接口获取互联网上与待标注图像数据相关的合法文本描述信息，作为辅助提示信息呈现给标注师。系统的整体设计如图 1 所示。

2 系统实现

为满足众包标注场景下多人数据标注师并发同步访问，

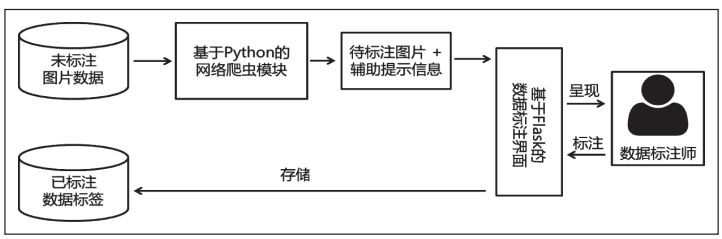

图 1　系统整体设计

标注平台的服务器部分采用 Nginx 实现。主要业务逻辑部分基于 Flask 框架实现，并采用 SQLite 数据库实现原始图像数据、标签数据和标注师用户数据的存储。为实现友好易用的用户界面，平台的前端基于 BootStrap 框架实现，并通过 HTML 和 CSS 控制页面风格的美观性与统一性。同时，利用 Python 的 requests 程序包实现爬虫模块，从互联网中获取合法的辅助提示信息。系统的整体框架如图 2 所示。

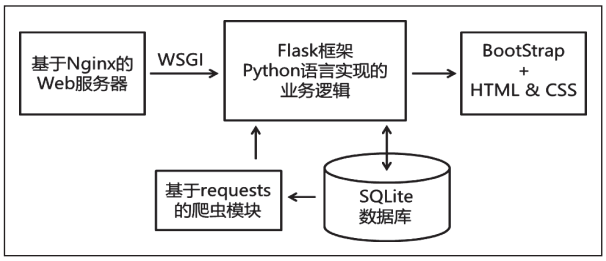

图 2　系统整体框架

如图 2 所示，本文所提出的可视化深度学习标注系统包含 5 个模块：基于 BootStrap 的前端表示层，基于 Flask 的核心业务逻辑，基于 Nginx 的 Web 服务器，基于 requests 库的网络爬虫，以及基于 SQLite 数据库的持久化模块。各模块的具体实现如下所述：

（1）基于 BootStrap 的前端表示层：该模块负责为数据标注师提供友好易用的可视化界面。每个具体页面通过 HTML 控制界面的布局，并通过 CSS 批量控制多个页面的风格与样式。同时，该模块负责处理数据标注师与页面中的交互，如提交数据、刷新界面等。控件的响应与刷新逻辑由基于 BootStrap 的 JavaScript 脚本实现。

（2）基于 Flask 的核心业务逻辑：该模块负责整体协调其他各模块的输入输出数据，实现标注平台网站的核心逻辑。在与前端表示层交互时，该模块根据所收到的请求查询和修改数据库中的数据，并访问 Web 服务器上的资源。同时，在数据标注的过程中，经过爬虫模块从互联网中爬取到的辅助标注信息，也由业务逻辑模块进行封装并通过前端表示层呈现给数据标注师。

（3）基于 Nginx 的 Web 服务器：相比于业务逻辑模块，该模块负责处理更底层的业务。如根据配置对不同的请求做出不同的转发，通过动静分离技术使整个系统具有更高的性能和效率。

（4）基于 SQLite 的数据库：该模块负责将原始数据，标注后的数据，以及用户的账号数据进行持久化，存储在服务器主机的磁盘中。在标注与查询过程中，该模块根据收到的请求对标注数据进行查询、修改、新增、删除等操作。

（5）爬虫模块：该模块基于 Python 的 requests 程序库实现。该模块将原始的图像数据进行处理，封装成网络数据包并调用第三方网站（如百度，谷歌，搜狗等）提供的程序接口，搜索和获取互联网中与待标注图像数据相关的文本信息，并以标准化的格式传送给业务逻辑层。该模块通过工厂设计模式进行封装，可方便地和更多的第三方程序库进行对接。

整套网站系统的开发与调试基于 PyCharm（Community Version）集成环境。软件代码调试无误，并经黑盒测试确认所有需求被实现后，部署在 Ubuntu Server 16.04 服务器主机上提供数据标注服务。

3 结论

本文论述了一个基于 Flask 网站开发框架和 Python 爬虫技术的深度学习数据标注平台的设计与实现。所论述的数据标注平台可以有效地利用基于 Python 的深度学习程序库扩充功能，并具有界面友好，易于扩展和维护的特性。同时，本文所述的数据标注平台可以通过网络爬虫技术向数据标注师提供有效的辅助信息，提升数据标签质量。本文的研究内容对基于深度学习的数据标注方法研究有积极意义。

参考文献

＊ [1] 聂震云 . 基于众包的数据标注系统 [D]. 北京交通大学 ,2014.

＊ [2] 蔡莉 , 王淑婷 , 等 . 数据标注研究综述 [J]. 软件学报 ,2020,Vol.31Issue(2):302-320.

＊ [3] 于观贞 , 陈颖 . 实体瘤病理数据集建设和数据标注质量控制专家意见 (2019)[J]. 第二军医大学学报 ,2019.

＊ [4] 崔爽 . 数据标注师 : 人工智能背后的人工力量 [J]. 科学中国人 ,2019.

＊ [5] 田红梅 , 周皎 , 袁志杰 , 等 . 基于 Python 的训练集数据标注修改方法研究 [J]. 科技创新导报 ,2020,000(004):208-209.

＊ [6] 杨佩军 . 众包数据标注质量的改善算法研究 [D]. 华东师范大学 ,2019.

＊ [7] 白雪丽 . 浅析基于 Python 爬虫技术的特性及应用 [J]. 山西科技 ,2018,v.33;No.186(02):58-60.

＊ [8] 管小卫 . 网络爬虫探讨及应用 [J]. 科技创新与应用 ,2020(27):178-179.

＊ [9] 徐昊 , 沈江明 . 面向网站群的主题爬虫研究 [J]. 软件导刊 ,2020,19(08):109-112.

＊ [10] 赵文杰 , 古荣龙 . 基于 Python 的网络爬虫技术 [J]. 河北农机 ,2020(08):65-66.