学号：2013012982

西北农林科技大学

2017 届本科生毕业论文（设计）

题目：基于 TensorFlow 的目标检测与识别

学院 (系)：	信息工程学院
专业年级：	计算机科学与技术 131 班
学生姓名：	贺文静
指导老师：	何进荣
合作指导老师：	
完成日期：	2016 年 6 月

基于 TensorFlow 的目标检测与识别

摘要:

关键词:; 计算机视觉; 深度学习;TensorFlow;

Research on visualization method of apple genome

Abstract: With the advent of the era of big data, various organisms omics data (such as genome,proteome, etc.) showing a trend of rapid growth. Through preliminary investigation found that a morethorough study and organisms (such as Arabidopsis, yeast, etc.) omics data, compared to Malus'smulti-omics data resources in recent years, was able to open access, the data distribution is morefragmented, lack of a unified and efficient management. In order to solve these problems, the establishment of a multi-omics malus Management System, Malus's multi-omics data integration of resources and effective management. The system uses Struts2 framework to MySQL database, Apache server, combined with JavaScript, Css, Html, Bootstrap and other languages and technology, (1) to achieve a malic 95,232 genes and 95,232 protein sequences, delete, query , modify the basic operation;(2) the use of BLAST sequence alignment tool, Malus realized gene and protein sequence similarity comparisons and screening; (3) in order to better understand the apple gene or protein function, the system also provides Arabidopsis Malus similarity search function in gene and protein sequences aspects of visualization in the genome-wide similarity metric, the site provides links to functional annotation to help users of the system by means of research more thorough gene or protein function in Arabidopsis, reasoning apple corresponding homologous gene or protein function. The Malus species researchers have some practical value

Keywords: Management System; Struts Framework; Apple Multiple Omics; Sequence Alignment

目 录

第一章　绪论

1.1　研究背景

1.1.1　人工智能的发展

无孔不入的人工智能，别应用在各个领域，比如谷歌传统的搜索和广告业务、无人驾驶汽车，以及医疗健康部门，苹果 Siri：为你解决问题的人工智能，应用了人工智能的百度外卖，代替人工顾问的智能应用，作为医疗辅助等。从互联网巨头到初创企业，都将人工智能作为发展的核心，如果关注这方面的新闻，应该看到过这样的一些新闻，” 百度以近 12 亿元剥离游戏业务”,” 央行成立金融科技委员会：用人工智能、云计算丰富监管手段”,” 苹果斥资 2 亿美元收购人工智能公司 Lattice “等。人工智能变得越来越重要，两会期间，讯飞科大的语音识别及人工智能产品展示时，赢得了一众喝彩。　人工智能不仅涉及到民用，也涉及国家各个核心战略领域。国家主席习近平 “一带一路 “中演讲。他表示，“一带一路” 建设要坚持创新驱动发展，加强在数字经济、人工智能、纳米技术、量子计算机等前沿领域的合作，推动大数据、云计算、智慧城市建设，连接成二十一世纪的数字丝绸之路。在这个人工智能到来的时代，发展人工智能变得越发重要。无论在国内国外，人工智能的发展都被寄予厚望。

1.1.2　深度学习的走红

随着人工智能的迅速走红，深度学习一词也迅速走红。人工智能是应用范畴的词汇，机器学习是一种实现人工智能的方法，深度学习是一种实现机器学习的方法，也是现有机器学习方法中，最奏效的一类。三者的关系见图 1-1，人工智能是最外面的，下来是机器学习，最里面的是深度学习。而计算机视觉是机器学习应用最成功的一个方面，发展最为迅猛的一个分支。计算机视觉在很多方面都有应用，最为大家熟知的有人脸识别，组织恶意软件，语音识别，无人驾驶汽车，在生物，医学方面也都有一定的应用，实际上，思科的一个评估显示，在 2016 年互联网上超过 85% 的信息都会是像素形式，我们进入 “多媒体” 时代，视觉的时代，图片和视频的时代。由于互联网作为信息的载体，以及大量（视觉）传感器引发了视觉信息的大爆炸。CS231a 深度学习与计算机视觉课程开篇，李菲菲教授讲计算机视觉（CV）她将这些视觉信息称为 “互联网中的暗物质”，并且举了 YouTube 的一个例子，我们没法对这么大量的数据进行浏览，标记，分类，索引等，但是使用计算机视觉技术能够对照片进行标签、分类、处理视频中的每一帧，自动截取出篮球比赛中—-比如说科比的一次精彩进球，我们现在面临着的问题就是，大量的数据，以及这些 “暗物质” 的挑战。目标识别与检测作为计算机视觉的一个应用，是很有应用价值的问题，而且是很多问题的根本解决之策。

1.1.3 TensorFlow 的出现

2015 年 11 月，谷歌宣布 TensorFlow 开源，由于其灵活的架构可以在一个或多个 CPU，GPU、桌面、服务器、以及移动设备上部署，还不用重新编写代码，其分布式计算的方法大大缩短了机器学习的训练时间，核心代码是 C++ 编写的简化了线上部署的复杂度，还有 Python、GO、Java 的接口，用户可以很容易的使用，可视化的 TensorBoard，极快的编译速度，并行计算模式等优点，使其作为一个开源软件库，刚开源第一个月就积累了 10000+ 的 star，而到现在 star 数已经到了 56939，是 GitHub 上最受欢迎的深度学习框架。在图形分类、音频处理、推荐系统和自然语言处理等场景下都有丰富的应用。最近流行的 Keras 框架底层默认使用 TensorFlow，著名的斯坦福 CS231n 课程使用 TensorFlow 作为授课和作业的编程语言，而且现在除了谷歌内部大规模使用之外，优步（Uber）、Twitter、京东、小米、FaceBook 等都在使用。TensorFlow 的 contrib.learn 模块提供了一个让开发人员从 scikit-learn 或 Keras 进入到 TensorFlow 的桥梁，避免了从一个框架转移到另一个框架的无措感。TensorFlow 的出现使更多对机器学习感兴趣的人可以去涉足这个领域，而不是因为电脑硬件的问题，机器学习训练时间过长的问题驻足。人工智能，机器学习，深度学习，TensorFlow 成为越来越受欢迎的话题，我在百度指数，以及 Google trends 上，分别输入了机器学习，深度学习，还有人工智能几个搜索词后，得到的搜索热度趋势图见图 1-2。通过这些数据得出的结论是人工智能，深度学习，正在变得越来重要，正在引起越来越多的重视。而在 Google trends 中看到 TensorFlow 搜索热度趋势图见图 1-3 。TensorFlow 的热度基本维持在 100，足以说明这是目前很受欢迎的一个话题。

1.2 深度学习

1.2.1

1.3 本文研究内容

1.4 本文内容安排

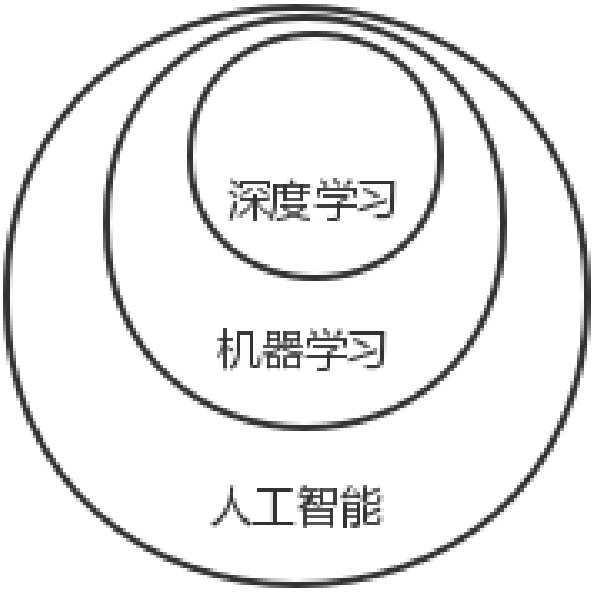

图 1.1:

图 1.2:

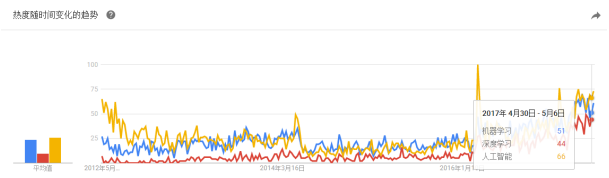

图 1.3: Google trends 搜索热度趋势图：词汇的热度按 0 100 分为 100 个等级，可以看出来三者都呈上升趋势，并且相当的受欢迎。

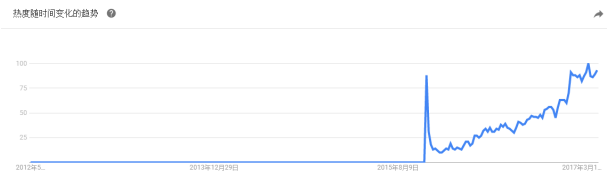

图 1.4:

图 1.5:

第二章　深度学习平台——TensorFlow

2.1　TensorFlow 简介

TensorFlow 在 2015 年 11 月由 Google 开放，从此，它已经成为 GitHub 上最受欢迎的机器学习库。TensorFlow——第二代分布式机器学习算法实现框架和部署系统，是基于使用 DisBelief 时的经验及训练大规模分布是神经网络的需求开发的。但是在某些基准上，TensorFlow 是 DistBelief 的两倍。可以方便地部署到各种平台，大大简化了真实场景中应用机器学习的难度。TensorFlow 的计算可以表示为有状态的数据流式图，使用数据流式图来规划计算流程，可以将计算映射到不同的硬件和操作系统平台。对于大规模的神经网络训练，TensorFlow 可以让用户简单地实现并行计算，同时使用不同的硬件资源进行训练，同步或异步地更新全局共享的模型参数和状态。TensorFlow 是相对高阶的机器学习库，用户可以方便的用它设计神经网络结构，支持自动求导，核心代码是用 C++ 编写的，简化了线上部署的复杂度，可以在手机和 CPU 这种内存紧张的设备上运行。除了 C++ 接口外还有 Python、Go、Java 等接口。可以部署在一台或多台 CPU、GPU 上，兼容多个平台，包括 Windows、Linux、Android 等。有 TF.Learn 和 TF.Slim 等上层组件可以帮助快速的设计新网络，并且兼容 Scikit-learn estimator 接口，同时 TensorFlow 不局限于神经网络，数据流式图支持非常自由的算法表达，只要可以将计算表达成计算图的形式，就可以使用 TensorFlow。另一个重要特点是它灵活的移植性，可以将同一份代码几乎不经过修改就轻松的部署到任意数量 CPU 或 GPU 的 PC、服务器或者移动设备上。还有一个优势就是极快的编译速度，还有功能强大的 TensorBoard，能可视化网络结构和训练过程，对于观察复杂的网络结构和监控长时间、大规模的训练很有帮助。

2.2　核心概念

2.2.1　计算图

UCSC Genome Browser 是由 University of California Santa Cruz (UCSC) 创立和维护的，该站点包含有人类、小鼠和大鼠等多个物种的基因组草图，并提供一系列的网页分析工具。站点用户可以通过它可靠和迅速地浏览基因组的任何一部分，并且同时可以得到与该部分有关的基因组注释信息，如已知基因，预测基因，表达序列标签，信使 RNA，CpG 岛，克隆组装间隙和重叠，染色体带型，小鼠同源性等。用户也可以因为教育或科研目的加上他们自己的注释信息。UCSC Genome Browser 目前应用相当广泛，比如 Ensembl 就是使用它的人类基因组序列草图为基础的。

2.2.2　可视化方式

JBrowse 用 track 的方式进行可视化，提供平滑的动态移动和缩放功能，也有导航和通道的选择。JBrowse 可以展示多种 track 视图，除基本视图外，还可以显示非翻译区、外显子、内含子结构等。

2.2.3　可视化内容

UCSC 基因组浏览器提供了多样化的注释数据集（称为“轨迹”并以图形方式呈现），包括 mRNA 比对，DNA 重复元件的映射，基因预测，基因表达数据，疾病关联数据（代表基因的关系疾病）和市售基因芯片（例如 Illumina 和 Agilent）的映射。显示的基本范例是在水平维度上显示基因组序列，并显示 mRNA 的位置，基因预测等的图形表示。沿着坐标轴的颜色块显示各种数据类型对齐的位置。在单个坐标轴上显示这种大量数据类型的能力使浏览器成为数据垂直整合的便利工具。UCSC 浏览器与其他基因组浏览器区分开来的一个独特而有用的功能是显示器的不断变化的性质。可以显示任何大小的序列，从单个 DNA 碱基到整个染色体（人类 chr1 = 2.45 亿碱基，Mb）和完整的注释轨迹。研究人员可以显示单个基因，单个外显子或整个染色体带，显示数十个或数百个基因以及许多注释的任意组合。方便的拖放功能允许用户选择基因组图像中的任何区域，并将其扩展到占据全屏。

2.2.4　系统架构

UCSC Genome Browser 的开发，起源于一小段应用于 C. elegans 基因预测拼接图谱的 C 语言脚本，后期通过不断扩充，才变成现在这样强大的一个分析工具。现在 UCSC 的主要开发语言是 Java/Python，后台数据库依赖于 mysql，而且提供 mysql 的公共接口，只要用户本地电脑装有 mysql 客户端，就可以通过 UCSC 提供的接口访问网站后台的数据库；对于前台要求，UCSC 可以较好地兼容 IE、Chrome、Firefox 等主流网络浏览器。UCSC 是完全开源的，用户可以下载到完整源码。

2.2.5　运行机理

2.2.6　优缺点

优点：JBrowse 属于新一代基因组浏览器，作为 GBrowse 的继任者，是基于最新的前端技术开发的。在 JBrowse 中，服务器端的负荷极大地降低，后台服务器只需要向浏览器客户端发送数据文件，将繁杂的计算工作从服务端脱离出来，大量计算工作被合理分配到了前端。同时，JBrowse 对 Cookies 技术也得到了很好的支持，可以有效记录用户的喜好。

缺点：JBrowse 把可视化主要工作放在了浏览器端，但其可视化方法仅是一些普通的 HTML 标签实现，造成可视化不友好等问题。同时浏览器在绘制图像时需要运行大量的 JavaScript 代码，而且目前主流浏览器对 HTML5 中新标签的支持不完善，造成用户体验不佳等问题。

第三章　基因组浏览器环境搭建与配置

3.1　GBrowse 搭建与配置

3.1.1　操作系统安装与配置

3.1.2　源码安装 Apache 与 MySQL

3.1.3　GBrowse 全局配置文件编写

3.1.4　GBrowse 苹果基因组配置文件编写

3.1.5　全局配置 Apache 与 GBrowse 进行模块关联

3.1.6　全局配置 MySQL 与 GBrowse 进行模块关联

3.2　JBrowse 搭建与配置

3.3　UCSC Genome Browser 搭建与配置

第四章　数据的处理与导入

4.1　系统要求

4.2　数据转储

第五章　苹果基因与蔷薇科植物基因可视化对比验证

5.1　拟南芥基因可视化

5.2　红桃基因可视化

5.3　验证实验结果

第六章　总结与展望

6.1　总结

6.2　展望

参考文献

[1] Lincoln D Stein. "*Using GBrowse 2.0 to visualize and share next-generation sequence data*". *Briefings in bioinformatics*, **2013**: bbt001.

[2] 周琳, 孔雷 and 赵方庆. "生物大数据可视化的现状及挑战". 科学通报 *(*中文版*)*, **2015**, *60*(5/6): 547–557.

[3] 夏艳. 水稻比较基因组和进化生物学数据库的构建研究 [mathesis], **2013**.

[4] 孙磊, 陈璇, 唐红 *et al.* "基于 *GBrowse* 的多源长非编码 *RNA* 数据可视化系统 □". **2017**.

[5] 孙秀丽. 植物基因表达数据库的构建及共表达分析研究 [phdthesis], **2013**.

[6] 张海川, 李杰 and 王亚东. "基于 *Web* 的基因组浏览器研究现状". *Progress in Biochemistry and Biophysics*, **2014**, *41*(11): 1182–1190.

[7] 杨锡南 and 孙啸. "生物信息学中基因数据可视化". 计算机与应用化学, **2001**, *18*(5): 403–410.

[8] 王蕊 and 胡德华. "生物信息学数据库研究文献引文与热点分析". *Chinese Journal of Bioinformatics*, **2014**, *12*(4).

致谢

时间如白驹过隙，四年的大学生活转眼就结束了。

在这短短的四年里，经历了许许多多的事情，有快乐有悲伤，也曾遇到挫折也曾彷徨，好在身旁有许许多多帮助我的人，首先感谢我的导师于建涛老师，于老师工作十分负责，非常关注我的毕业设计进度并且在遇到困难的时候给予我思路和方向，于老师严谨细致的工作作风一直伴随着我整个毕业设计的过程，时刻激励着我对整个毕业设计的改进和优化。在此，我非常感谢于老师对我的毕业设计和论文的帮助和支持。

感谢我的舍友们，在这四年里是你们一直陪伴在我的身边，在我遇到困难的时候伸出援手帮助我度过难关。

感谢我的爸爸妈妈，无论什么时候都爸爸妈妈在背后默默支持我、鼓励我，让我顺利的完成了学业，并将继续支持我出国深造，是你们的支持给予了我前进的动力，让我有力量继续走下去。

现在论文即将完成，心情很是激动，回想起从一开始的选题到现在论文的完成，我经历了许许多多，在这期间也有许许多多的人给予我无私的帮助，在此请接受我最诚挚的感谢。