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

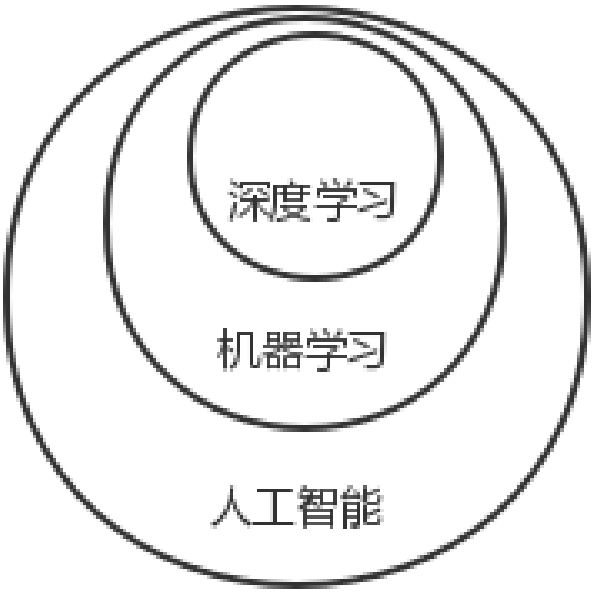

图 1-1

图 1-2

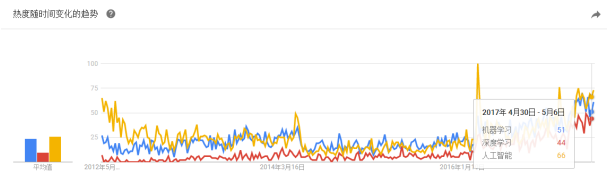

图 1-3 Google trends 搜索热度趋势图：词汇的热度按 0 100 分为 100 个等级，可以看出来三者都呈上升趋势，并且相当的受欢迎。

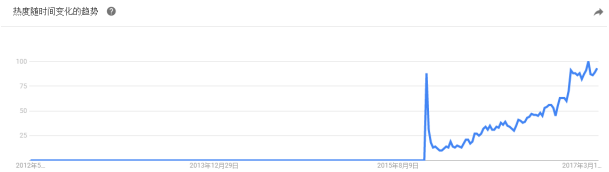

图 1-4

图 1-5

第二章　深度学习平台——TensorFlow

2.1　TensorFlow 简介

2015 年 11 月 Google 在 Github 开放 TensorFlow 的源代码，从此，它已经成为 GitHub 上最受欢迎的机器学习库。TensorFlow 是第二代分布式机器学习算法实现框架及部署系统，是基于使用 DistBelief 时的经验及训练大规模分布是神经网络的需求开发的。但是在某部分基准上，TensorFlow 的性能是 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 Java、Python、Go 等接口。TensorFlow 可以部署在一台或多台 CPU、GPU 的机器上，兼容多种平台，包括 Android、Windows、Linux 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 不仅仅局限于神经网络和数据流式图并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。另一个重要特点是可以灵活的移植性，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 上、服务器或者移动设备上。还有一个优势就是编译速度极快，在创建新的网络结构时，Theano 通常需要很长时间的进行编译，因此尝试新模型需要付出较大的代价，而 TensorFlow 不存在这样的问题。而且还有功能强大的 TensorBoard，能够将网络结构和训练过程进行可视化，对于研究者记录和观察复杂的网络结构和监控长时间、大规模的训练有很大的帮助。TensorFlow 不仅支持常见的神经网络结构外，而且还支持深度强化学习甚至支持密集的科学计算。TensorFlow 之前版本不支持 symbolic loop，需要使用 Python 循环解决这些问题，从而带来无法进行图编译优化等一系列问题，但最近新发展的 XLA 开始对 JIT 及 AOT 进行支持，另外其使用 bucketing trick 也可以较高效地实现循环神经网络。TensorFlow 的一个不占优势的方面，可能对于计算图必须构建为静态图，这使得很多计算实现困难。

2.2　核心概念

2.2.1　计算图

计算图是 TensorFlow 中最基本的概念之一，在 TensorFlow 中的所有计算都会被转化为计算图上的节点。在 TensorFlow 中计算可以表示为一个有向图（directed graph）或者称为计算图（computation

graph）。TensorFlow 是一个通过计算图的形式来表述计算的编程系统，其中每一个运算操作（operation）将作为一个节点（node），节点与节点之间的连接成为边（edge）。计算图中每一个节点可以有任意多个输入和任意多个输出，节点可以算是运算操作的实列化。在计算图的边中流动（flow）的数据被称为张量（Tensor），这个也是 Tensoflow 名字的由来。没有数据流动的边被称为依赖控制。计算图描述了张量数据的计算流程，负责维护和更新状态，对分支进行条件控制和循环操作。如果机器上有超过一个可用的 GPU, 除第一个外的其它 GPU 默认是不参与计算的. 为了让 TensorFlow 使用这些 GPU, 你必须将 op 明确指派给它们执行. with...Device 语句用来指派特定的 CPU 或 GPU 执行操作

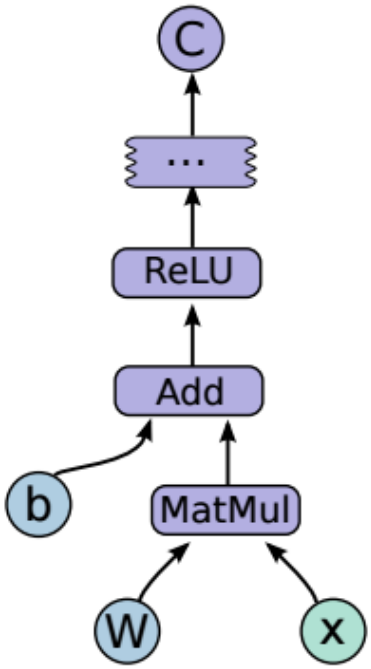

图 2-1 计算图示例

2.2.2 运行模型

Tensorflow 的运行环境使用的是 Session。用户通过使用 TensorFlow 交互式接口 Session 进行操作。TensorFlow 程序运行时的所有资源由 Session 管理。当所有计算完成之后需要关闭会话来帮助系统回收资源，否则就可能出现资源泄漏的问题。TensorFlow 中使用 Session 的模式一般有两种，第一种模式需要明确调用会话生成函数和关闭会话函数，第二种模式是通过 Python 的上下文管理器来使用会话，这种模式为了解决异常退出时资源释放的问题，TensorFlow 的会话不会自动生成默认的会话，需要手动指定，当默认的会话被指定之后可以通过 tf.Tensor.eval() 函数计算一个张量的取值，在交互式环境下（比如 Python 脚本等）通过设置默认会话的方式来获取张量的取值更加方便，函数 tf.InteractiveSession（）在交互式环境下直接构建默认会话。使用这个函数会自动将生成的会话注册为默认会话。可以省去将产生的会话注册为默认会话的过程，不论采用哪种方法都可以通过 ConfigProto Protocol Buffer 来配置生成需要生成的会话，通过 ConfigProto 可以配置类似并行的线

程数，GPU 分配策略，运算超时时间等参数。在这些参数中最常使用的是 allow_soft_placement= 和 log_device_placement 两个参数其都是布尔型的参数。在实现上, TensorFlow 将图形的定义转换为分布式执行的操作，是为了充分使用可用的计算资源 (如 CPU 或 GPU)。通常来说不需要显式指定使用 CPU 还是 GPU, TensorFlow 能主动检测。如果检测到 GPU, TensorFlow 会使用找到的第一个 GPU 来执行操作。

2.2.3 数据模型

张量是 TensorFlow 管理数据的形式，可以被简单的理解为多维数组。其中零阶张量表示标量（scalar），也就是一个数；第一阶张量为向量（vector），也就是一个一维数组；第 n 阶张量可以理解为一个 n 维数组。在张量中并没有真正存储数字，它存储的是如何计算得出这些数字的过程。TensorFlow 计算结果不是一个具体的数字，而是一个张量的结构，一个张量主要有三个属性：名字（name），维度（shape），类型（type）第一个属性名字不仅仅是张量的唯一标识符，还给出了这个张量是如何计算出来的，张量与计算图上节点代表的计算结果是相对应的，张量的命名能够由“node：src_output”的方式给出，其中节点的名称为 node，src_output 表示当前张量是来自节点的第几个输出。第二个属性是张量的维度（shape）描述了一个张量的维度信息。第三个属性是类型（type）每一个张量会有一个唯一的类型，TensorFlow 会对参与运算的所有张量进行类型检查，当发现类型不匹配时会报错。TensorFlow 支持多种不同的类型，主要包括了实数（tf.float16 tf.float32,tf.float64)、整数（tf.int16,tf.int32,tf.int64,tf.int8,tf.uint16,tf.uint8）、布尔型（tf.bool）, 复数型（tf.complex64,tf.complex128）等。对于张量的使用，可以分为两大类。第一类是对中间计算结果的引用。当一个九三包含很多中间结果时，使用张量可以大大提高代码的可读性。第二类是当计算图构造完成之后，张量可以用来获得计算结果，结合会话使用 tf.Session().run(result) 就可以得到真实的数字.

2.2.4 变量

在 TensorFlow 中，变量的作用就是保存和更新神经网络的参数。变量包含张量 (Tensor) 存放于内存的缓存区。建模时它们需要被明确地初始化，模型训练后它们必须被存储到磁盘。这些变量的值可在之后模型训练和分析是被加载。TensorFlow 提供了一个通过变量名来创建或者获取一个变量的机制。通过这个机制，在不同的函数中可以直接通过变量的名字来使用变量，而不需要将变量通过参数的形式到处传递，这个机制主要通过 tf.get_variable() 和 tf.variable_scope() 函数实现的。

2.3 Tensorflow 环境搭建

Tensorflow 对各种主流的操作系统的支持都比较完善，本文将使用 Tensorflow 对 Python 提供的 API 进行研究学习。Tensorflow 支持的 Python 环境有 Python2.7 和 Python3.5，目前最新版本的 Tensorflow 支持 Python3.5。Tensorflow 环境的搭建可以分为两类，第一种是使用官方提供的发行进行安装使用，第二种是下载 Tensorflow 的源代码进行编译安装使用。本文将以第一种方式在 Linux 和 Windows 系统环境下分别使用 Pip 和 Anaconda 进行环境搭建。

2.3.1　Pip 下环境搭建

Pip 是一个以 Python 计算机程序语言写成的软件包管理系统，其可以安装和管理软件包。通过 Pip 可以安装已经打包好的 TensorFlow 以及 TensorFlow 所依赖的其他 Python 扩展包。环境需保证操作系统上有 Python 的环境，首先，安装 Python 的 Pip 包，然后通过 Pip 命令针对不同版本的 Python 安装 CPU 或者 GPU 版本的 Tensorflow。如下安装过程的详细代码。

```
sudo apt-get install  python-pip python-dev
pip  install  tensorflow  # Python 2.7; CPU support (no GPU support)
pip3  install  tensorflow # Python 3.n; CPU support (no GPU support)
pip  install  tensorflow-gpu # Python 2.7;  GPU support
pip3  install  tensorflow-gpu # Python 3.n; GPU support
```

2.3.2　Anaconda 下环境搭建

Anaconda 是由 Python 开发的领先的开放数据科学平台。包括超过 100 种最受欢迎的数据科学 Python，R 和 Scala 软件包。内置了大量的 Python 经常用的库，包括 SciPy、Numpy、Scikit-learn、Pandas 等。首先，需要安装用于科学计算的 Anaconda 环境，其次创建 Python 不同版本的环境，最后通过使用 Python 的包管理工具 Pip 安装 TensorFlow 的环境。安装过程的详细代码如下。图 2-2 安装成功后测试图。

```
Anaconda3-4.3.1-Windows-x86_64.exe #Windows Install
bash Anaconda3-4.3.1-Linux-x86_64.sh #Linux Install
conda create --name python35 python=3.5 # create environment in Linux and Windows
source activate  Python35 # Linux
activate  Python35 #Windows
pip  install  tensorflow # install  TensorFlow
```

```
(python35) C:\Users\hewenjing>python
Python 3.5.3 |Continuum Analytics, Inc.| (default, Feb 22 2017, 21:28:42) [MSC v
.1900 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> import tensorflow as tf
>>> print(tf.__version__)
1.1.0
```

图 2-2 TensorFlow 安装成功

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。