**远程科研项目报告**

**——手写体数字识别**

贺英豪

2019年10月11日

1. **研究背景**

随着大数据技术的发展，对大数据的相应处理技术就显得十分重要，这次研究，通过对使用docker的容器部署MNIST手写数字识别模型来识别手写数字，并在Cassandra中展示结果这一过程来实现对大数据存储以及应用等方面的学习，进而对大数据的相关知识有相应的了解。

* 1. **Docker**

Docker 是一个开源的应用容器引擎，让开发者可以打包他们的应用以及依赖包到一个可移植的容器中,然后发布到任何流行的Linux机器上。Docker最大的优势在于其效率，当改变其中的镜像时，只需要改变其变化的部分，而不需要重建整个Docker镜像。本次实验实将使用Docker的容器来进行。

可以通过pip install docker命令下载docker，下载完成后通过docker --version参看其版本看是否安装成功。

如果国内网速慢可以在后面加上下列代码使用清华的源（下同）

-i https://pypi.tuna.tsinghua.edu.cn/simple

* 1. **Mnist与TensorFlow**

简单来说MNIST数据集是一个手写体数据集，这个训练集是由60000个用例组成。通过这个训练集我们可以实现对手写体数字的识别。

TensorFlow是一个基于数据流编程（dataflow programming）的符号数学系统，被广泛应用于各类机器学习（machine learning）算法的编程实现。本次将使用TensorFlow实现对mnist模型的保存于调用。

通过pip install tensorflow

conda install -c conda-forge tensorflow

完成TensorFlow的下载。

**1.3 Cassandra**

Cassandra是一套开源分布式NoSQL数据库系统，具有灵活，可扩展性等诸多优点，本项目将使用Cassandra来进行结果的记录与存储。

通过dokcer pull cassandra来讲Cassandra容器拉到本地。这个过程比较慢（国内网络问题），如果不成功也可以使用如下命令：

docker pull [registry.docker-cn.com/library/cassandra:latest](http://registry.docker-cn.com/library/cassandra:latest" \t "https://mail.qq.com/cgi-bin/_blank)

1. **具体实现**

**2.1 虚拟环境配置**

由于docker与Windows的兼容问题，本次项目我将在Ubuntu系统上来完成，虽然虚拟环境的配置与项目本身关系不大，但是通过配置好虚拟环境，更有利于Linux系统上的后续操作。

**1安装：pip install virtualenv**

**2创建虚拟环境：virtualenv hyh（环境名） -p /usr/bin/python3.7**

**3进入虚拟环境：source bin/activate**

创建好后下次进入只需要 workon hyh即可。

**2.2 Mnist模型保存**

通过在mnis\_deep.py或者mnist\_softmax.py中加入如下代码实现文件的保存：

saver = tf.train.Saver()

with tf.Session() as sess:

Init = tf.global\_varianles\_initializer()

sess run init

#在""中填入保存路径

save\_path = saver.save(sess, "./modle.ckpt")

保存成功后会生成checkpoint、modle.ckpt.meta、modle.ckpt.data-00000-of-00001、modle.ckpt.inde、四个文件。需要调用时可通过saver.restore(sess, "./modle.ckpt")调用。

在训练时，官方给出的mnist训练文件中，定义各种key Variable的时候都加了：with tf name\_scope(＇＇)

这使得这些key variable的路径前都有了＇＇中给出的内容，发生了变化导致运行程序的时候报错找不到key variable。因此我删除了官方给的mnist\_deep中的这些定义，路径就不会发生错误，这样就可以找到这些变量。

这个问题困扰了我很久，我一直不明白训练好的训练集怎么会出现找不到的情况，后来通过自己编写了一个check.py的程序寻找这些变量的路径，我才发现路径不统一的问题，这个问题确实，麻烦，因为好几个同名变量在几个不同路径下，调用时很麻烦不如删除路径，都放在一个路径。下面是查询这些变量的代码：

from tensorflow.contrib.framework.python.framework import checkpoint\_utils

var\_list = checkpoint\_utils.list\_variables("./model.ckpt")

for v in var\_list:

print(v)

由于电脑配置问题我的训练过程运行十分缓慢，20000步对我电脑来说太多，因此为了快速得到模型投入使用，而且本实验对识别正确率没要求，就减少了训练步数，因此最终暂时的时候有个数字没能正确识别，这个问题只需要有点耐心多训练几步多等一会就能解决。

**2.3 Cassandra配置**

通过docker network create some-network创建一个网络

在通过docker run --name hyh(名字)-cassandra --network some-network -d cassandra:latest运行

最后docker run --name hyh-cassandra -p 9042:9042 -d cassandra:latest将其映射到9042端口。

如果要在python中使用Cassandra还需要通过如下命令下载相应的驱动程序：

pip install cassandra-driver

这一步比较容易，除了Cassandra的镜像可能不太容易pull下来（网络原因），剩下的只需要注意映射到正确的端口就行。

**2.4 程序运行**

**2.4.1本地运行**

先在最终的app.py文件路径下打开终端，通过workon hyh进入之前生成好的虚拟环境，然后直接python app.py就行。运行好程序后，通过docker start的命令启动我们之前生成好的容器等会用于运行Cassandra。

之后在开启另一个终端，通过curl localhost:7000/mnist(自己定义的地址) -F ＂file=@path to file＂上传文件。

这一步首先要注意我们用的是POST方法，网页端直接打开是不行的因为网页无法使用POST方法。因此需要使用curl命令上传。其次如果上传图片报错图片大小问题是因为目前的程序暂时只能识别28\*28的图片，解决这个问题最简单的方法是打开PS直接压缩一下图片，也可以编写程序，通过OpenCV的cv2.resize命令调节大小。调节好图片就可正常上传识别了。我也上传了我做好的0-9的28\*28图片方便后来的同学直接使用。

识别成功后，就可以通过docker run it 容器名 bash进入之前的容器运行Cassandra。通过use mnist\_data1（自己的目录名）进入我自己的目录，在通过select\*from mytable就能看到数字，时间戳等信息。

**2.4.2 容器运行**

这一步中容器一直说找不到某个文件，应该是虚拟环境问题没弄好导致找不到一直报错，又因为有托福考试没空处理，容器运行只能作罢，不过大概思路我还是明白的，因此就讲一下大体的思路。

首先通过docker build --tag=final(名字) .生成一个docker image。这一步注意要讲要下载的东西全部写入requirements，同时建议在dockerfile中的pip install命令最后加上上面给出的清华镜像不然生成过程会因为下载速度过慢而timeout报错。

生成好之后就可以运行了，通过如下命令

docker run -p 4000:7000 final

将镜像映射到4000端口，避免因7000端口被之前占用而报错。没占用直接docker run也是可以的。

之后的步骤就与之前本地运行相同了。因此就不过多说明了。

1. **心得体会**

这次的项目终端来说其实不算难，但是在进行过程中真的遇到了好多好多的小问题导致进展的不顺利，首先最大的问题就是网络，Linux系统使用的各种软件源都来自国外，这导致很多东西的下载速度非常缓慢，甚至会下不下来。其次各种安装包的版本也是个问题，之前下载了TensorFlow最新版本结果导致无法运行还很难找到错误，最后卸载新的版本下载旧的才得以解决问题。这种版本以及虚拟环境的问题报错真的耽误了很多时间。

不过这次的收获也很多，虽然项目不复杂，但是老师的课程确从整个大数据的结构框架入手，让我对大数据的整体框架有了一个系统性的了解，同时也对大数据不同层面所使用的软件，所应当掌握的技术有了初步的认识。这对于我来说确实是一种全新的体验，之从来没接触过大数据的我也开始对这方面产生了兴趣，将来或许可以考虑选择这一个方向。

最后还得感谢老师这么多天的细心指导，很多问题基本是在老师的指导下才能顺利解决。同时老师也为我扩宽了知识面与视野，让我对大数据方向有了进一步的认识。