AI数据训练方面的学习进展

1. **数据的收集**
   1. **自行下载数据集**

在Kaggle上搜索关键词“food types”，筛选下载了3个数据集。



432700\_821742\_bundle\_archive：大小1G，下分evaluation/training/validation三个子目录，每个子目录下是存有数量不等.jpg文件的11个食物种类目录。

1864\_33884\_bundle\_archive和8544\_11959\_bundle\_archive：大小分别5G和10G，把食物分类到了101个子目录（两数据集都有相同结构的101个目录）。未区分training/validation，如选用这两个数据集需自行按比例划分训练/测试数据集。1864\_33884\_bundle\_archive还包含数个扩展名为.h5的文件（见下图），在Python中加载失败，提示“ValueError: No model found in config file.”



经与老师沟通，选用11分类的432700\_821742\_bundle\_archive数据集，进行了训练。

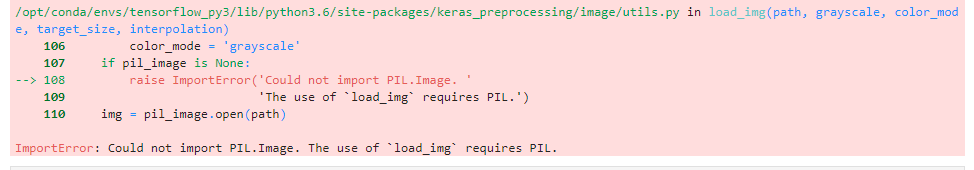
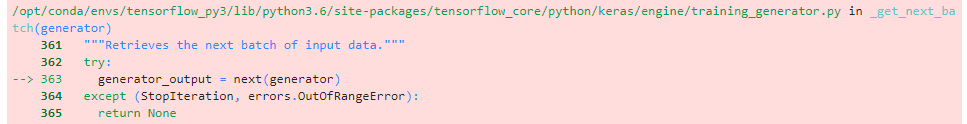
1. **图像识别群里提供的水果蔬菜数据集**

老师提供了水果蔬菜数据集的Github链接，由于本队目标是总体食物分类，蔬菜水果只是子集，所以没有下载研究。

1. **AI模型训练**

**在任务书脚本基础上，修改各类参数。**

* 1. **一些发现（部分心得亦可用于猫狗识别模型改进）**
     1. 由二分类改为11分类，flow\_from\_directory()函数应设置class\_mode = “categorical”而不是猫狗识别中的“binary”；
     2. 修改训练次数epochs对模型文件大小没有影响；
     3. 修改batch\_size，由128改为32后，训练结果模型由121M降至81M；
     4. 猜测 batch\_size 一定范围内调小有助于提高识别准确率，但因每次训练耗时较多而未及充分验证；
     5. shuffle设为False可用于多次测试时取得同一批影像以便统一比较标准；
     6. 训练一定轮次后，training accuracy不断上升的同时val\_accuracy反而下降，猜测是过拟合所致：在模型定义中引入Dropout()参数后该现象缓解，即val\_accuracy不再呈下降趋势，但升到一定数值（0.5）后小范围波动也不再上升。
  2. **碰到的问题**
     1. CNN卷积神经网络建立起来了（food11\_30\_cnn.h5），但ResNet模型遇阻；
     2. 腾讯云智能钛环境（目前都是在自己的无GPU电脑上训练，比较耗时）
        1. 按照使用指南操作了一下，很不熟练，容易操作失误反而耗时更久；
        2. 在智能钛云端，选定资源为TI.GN10X\_2XLARGE40.1Xv100（32G显存8核CPU），Launcher选定Notebook下conda\_tensorflow\_py3环境，试跑任务书的猫狗模型，到运行fit\_generator()脚本时，有如下报错：



（其实在前面执行画图的代码已有报错，但因为那不是必要代码而已忽略之）

上网搜了一些资料，似乎import pillow可解决该问题。然而Python环境下无法自行pip安装；如New Launcher选择Terminal，可pip install，但亦需安装TensorFlow；然后从Terminal进入Python环境后，执行相关指令时提示未发现可用GPU环境。故模型训练目前全在本地CPU上进行，耗时较长。