实训大作业： Kaggle 图像识别赛题——植物幼苗分类

1. 题目背景：

   你能区分杂草和作物幼苗吗？

   有效地做到这一点的能力意味着更好的作物产量和更好的环境管理。某大学信号处理小组与南丹麦大学合作，最近发布了一个数据集，其中包含属于 12 类物种的大约 960 种独特植物在几个生长阶段的图像。很多时候很难从幼苗中分辨出是哪种植物。在这次实训的最后，我们将制作一个 ML 模型，它可以帮助我们识别幼苗，并告诉我们它是哪一种植物。

   注：因为这个案例是赛题，所以测试集的标签不予公布，因此本次大作业只在训练集上完成即可。

2. 实验的重难点：
   1) 图像特征提取；
   2) 深度卷积神经网络结构搭建；
   3) 可视化分析；
   4) 模型调优；
3. 实验环境：

   Python、Keras(或者 tensorflow\pytorch)、matplotlib、seaborn 等；
4. 作业要求：
   1) 读取给定 train 集数据样本可视化几个实例；
   2) 数据样本进行特征提取；
   3) 构建样本特征向量空间；
   4) 样本可视化分析，例如样本分布、特征分布、特征相关性等（越多越好）；
   5) 搭建深度卷积神经网络模型；
   6) 将 train 数据划分训练集测试集（验证集）；
   7) 训练、评价、测试模型性能，绘制 loss acc 曲线（及多种可视化分析性能的图）；
   8) 随机从各类样本抽取一张图片存放到文件夹（无需编程、手动即可）；
   9) 将模型保存到本地，编写程序读取网络模型，对文件夹中随机选中的图片进行预测，
   10) 编写一个界面（可以是 QT、Tkinter 这种 Win 程序形式、也可以是 flask\django 等 web 形式），一个选取图片的按钮，一个原始图片可视化展示区域、一个预测结果展示区域，可以实现选择一张图片加载进来并预测文件夹中的每一张图（15 分）；
5. 评分标准（满分）：

   (1)**10 分**、(2)**10 分**、(3)**5 分**、(4)**20 分**、(5)**10 分**、(6)**5 分**、(7)**20 分**、(9)**5 分**、(10)**15 分**
6. 提交材料：
   1) 源代码、可视化分析结果、算法简报；
   2) 命名规范：第**组_大作业_ Kaggle 图像识别赛题——植物幼苗分类；
   3) 邮箱不变；
7. 验收流程：

   验收标准：以小组为单位进行逐步骤讲解演示，大作业成绩为 70%作业成绩+30%验收成绩，组内各步骤讲解分工由组内自定，不强制要求每个人都讲解演示，但参与演示的同学才有验收成绩。

   验收形式：通过线上会议现场演示或录制视频的形式，具体情况提前通知。

附录

数据样本说明：

- train.csv - 训练集，植物种类按文件夹分类；
- test.csv - 测试集，比赛时需要预测每个图像的品种（本次作业不适用）；
- sample_submission.csv – 比赛时提交的预测结果数据表（本次作业不适用）；

数据样本类别说明：

i. Black grass 苜蓿；

ii. Charlock 野芥子；

iii. Cleavers；猪殃殃；

iv. Common Chickweed 长毛箐姑草；

v. Common wheat 普通小麦

vi. Fat Hen 藜；

vii. Loose Silky-bent ；

viii. Maize 玉米

ix. Scentless Mayweed 淡甘菊；

x. Shepherds Purse 芥菜；

xi. Small-flowered Cranesbill 天竺葵；

xii. Sugar beet 甜菜；