模式识别与机器学习大作业：细胞图像识别与分割

2020. 5. 12

1. 问题描述

随着显微成像与荧光成像技术的发展，细胞图像能够为人们提供的信息越来越丰富，例如细胞的形态、数量、分布以及基因表达量等等。通过机器学习方法对细胞图像进行图像分割，智能化地提取细胞图像中的细胞及其所包含的信息，在医疗诊断、生物学研究等领域有着重要的应用价值。在本项目中，同学们将对细胞图像分割进行一个初步的探索，结合课内外所学的模式识别与机器学习方法，设计分割算法，进行细胞图像中每个细胞的实例分割(Instance Segementation)。本次大作业共计100分，具体任务要求及分数分配如下：

1. 结合你所学的模式识别与机器学习算法，在数据集1（见附件- dataset1）上进行训练与测试，提交你认为分割效果最佳的一组测试结果，数据说明、结果评价标准及提交文件格式见下文（40分）。
2. 撰写实验报告，对（1）中的数据处理流程、算法原理、实现过程、实验结果、模型性能等方面进行说明和分析，请注意报告的整体质量（60分）。
3. 【选做】对数据集2（见附件- dataset2）进行训练与测试，提交你认为分割效果最佳的一组测试结果。与（1）不同的是，数据集2中的图像只有少量标注，因此同学们可能需要考虑如何扩充数据集进行训练。（最高附加10分，但大作业整体评分不高于100分）。
4. 数据说明
   1. **文件命名格式**

作业提供了两个数据集，分别对应问题（1）和（3），每个数据集的结构与文件命名规则如下所示：

数据集L

------ train （数据集L对应的训练集细胞图像文件夹）

------ xxx.tif（对应训练集细胞图像xxx）

------ train\_GT/SEG（数据集L对应的训练集标注图片文件夹）

------ man\_segxxx.tif（对应训练集细胞图像xxx.tif的标注）

------ test （数据集L对应的测试集细胞图像文件夹）

------ xxx.tif（对应测试集细胞图像xxx）

* 1. **结果提交格式**

数据集L

------ test\_RES（你的分割结果图片文件夹）

------ maskxxx.tif（对应测试集细胞图像xxx的分割结果）

将文件夹“数据集L”打包成“数据集L.zip”格式，并在网页对应的入口提交即可。需注意的是，提交结果文件请保存为uint16格式，并请勿用中文命名。

* 1. **细胞图像数据**

我们的细胞图像来自于一段细胞影像的部分帧，每一张图像都存储为16位的tiff格式，因此像素的范围是0~65535而非常见的0~255。请注意，原图像较暗，因此若直接用看图软件打开无法看到图中的细胞，同学们需要对原图像的像素范围进行合适的调整。

* 1. **细胞图像标注**

每张图像的标注结果是一张与原图像大小相同的16位tiff图像，其中背景区域的像素值对应为0，细胞区域的像素值为特定的正整数，即假如原图像中共有N个细胞，则标注图像中每个细胞区域所对应的像素值分别为1~N。同理，提交的分割结果格式与标注相同。

考虑到数据集2中标注图片较少，我们提供了一个标注软件供大家使用，同学们可以使用其扩充训练集。软件使用说明见附件-label-software。

* 1. **结果提交**

我们搭建了专门的评测网站供同学们在线测试。输出分割结果后， 将结果文件“数据集L.zip“上传至评测网站（网站将稍后发布，请同学们关注网络学堂及微信群消息），后台进行评测后，会在网页返回对应数据集的评分及当前排名前5的评分。**每名同学限制提交10次有效结果(即如果格式不对出现的无法评分等情况不计入次数)**。

* 1. **代码样例**

基于问题（1）中使用到的数据集1，我们提供了一段简单的分割和结果输出代码示例，供大家参考，（见附件- dataset1\_demo.py）。

* 1. **保留测试集**

为保证公平性，我们将保留测试集中一部分的数据，用于再次验证模型效果。保留数据与公开数据部分是对全部测试集的随机划分。理论上，保留部分与公开部分测试集上的分割效果不应有较大差别。这一部分不计入排名得分，但是如果有同学模型在保留测试集与公开的测试集上效果有很大差异，需要同学通过说明训练测试过程等方式，证明方法的有效性。

1. 评分标准
2. **评分细则**

本次大作业实验部分的评分（即问题（1）与（3）的评分），将以**测试集评价指标排名的高低**给出。测试集评分最高的对应为满分，最低的对应为0分，中间的同学将按照梯度进行插空。具体的方法以下例为说明：

假如共6名同学，对于第（1）问的评测结果（满分40分）分别为0.2，0.5，0.5，0.75，0.9，1，那么得分就有5个梯度（评测结果相同的合并为一个梯度，精确到小数点后三位）：0分，10分，20分，30分，40分。然后依次将得分匹配到各个分数。如果出现了6名同学评测结果为0.2，0.9，0.9，0.9，0.9，1，那么将只有三个梯度：0分，20分，40分。

关于第（3）问10分的评分方式：参照（2）的要求提供详细的实验报告和文档，在此基础上按照预测性能进行评分，评分细则与（1）相同。只提交预测结果，不提交详细文档说明则不得分。

其余得分为实验报告的评分，将根据报告整体质量进行评判。

1. **评价指标**

在本项目中，我们采用匹配细胞间的Jaccard相似度来评价分割结果。对于真实标注图片中的一个细胞R与分割结果中的一个细胞S，我们称R与S相互匹配，当且仅当：

其中表示细胞区域在图片中对应的像素个数。对于匹配的标注细胞R和分割结果S，Jaccard相似度计算方式如下：

可以知道，Jaccard相似度取值范围在之间，其中1代表完全匹配，0代表完全不匹配。多个细胞的Jaccard相似度取平均为最终评分。

1. 提交格式

最终的作业请以压缩包的形式提交至网络学堂，压缩包中需包含实验报告（pdf格式）、所有源代码（例如\*.py格式）、代码环境配置、模型说明以及模型参数及权重，运行说明的README文档以及测试结果文件夹。其中测试结果文件夹及其中图片的格式与命名规则请参考**2.1**至**2.2**。**最终成绩会对提交代码进行再测试，验证后给出**。

1. 注意事项
2. 请仔细阅读大作业文档，弄清任务要求、数据内容以及提交格式，若完成大作业问题中遇到任何问题，请及时与助教联系。
3. 完成大作业时，同学们可采用任何合适的机器学习算法进行数据处理与模型训练，算法的实现允许调用第三方包及深度学习框架。
4. 请按照要求的提交格式提交最终的分割结果，否则可能出现0分的情况。
5. 网页提交结果与最终提交的压缩包中**请不要上传数据集**，只按要求提交测试集分割结果即可。
6. 请按照问题（2）中的要求认真撰写项目报告，实验报告占据大作业60%的分数，所以无论结果如何一定要提交报告，失败也是一次有益地尝试（认真撰写报告也可以拿到及格分）。

最后，祝同学们身体健康，学业顺利！