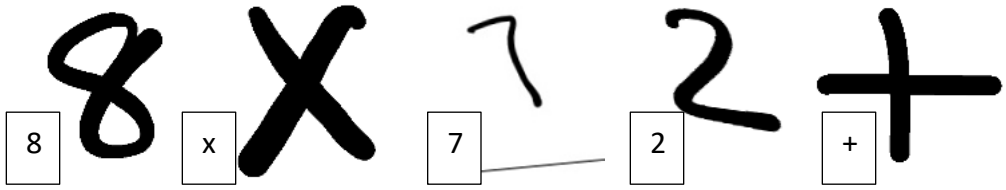
生物大数据分析第三次作业

任务 1：利用 CNN 算法完成 final_symbols_split_ttv 数据集的分类

数据集介绍：
该数据集包含 50000 多张图片，标签分别是 0-9 的数字和四则运算符号共 14 类。图片数据均是分辨率为 150*150 的灰度图像，目标是搭建卷积神经网络模型实现多分类问题。



数据集已经被划分为 train、test 和 val 三个子数据集，分别用作训练、测试和验证。在每个子数据集文件夹中，使用子文件夹存储了每个类别的图像，共 14 个类别。

提示：可以使用 PyTorch 的 ImageFolder 数据集类来加载数据集。在 ImageFolder 加载数据集后，文件夹的名称将被视为标签，并为每个图像分配相应的标签。例如：
train_set = torchvision.datasets.ImageFolder(root='./ final_symbols_split_ttv/train')
这样直接加载 train 子数据集就会自动为每个图像分配其所在文件夹的对应名称的标签。再使用 torch.utils.data.DataLoader 加载 train_set 就可以获得一个提供 image, label 的迭代器，可以直接用于后面模型的训练和测试。

作业要求：
1、使用 MLP 和 CNN 两种深度学习算法完成。MLP 多层感知机，即最基础的全连接神经网络。在搭建 **CNN 模型**完成任务时，要求实现 **3 层卷积层**，并使用**随机梯度下降**优化和**交叉熵**损失函数，其他的超参数可以自己调参和探索。MLP 模型不做限制，但模型体量应和 CNN 相当，用以后续比较结果。
2、代码应当包括：数据集加载和数据预处理、模型搭建、定义优化器和损失函数并训练模型、用训练好的模型进行预测。
3、比较和讨论你的算法性能、分类效果等，结合任务说一说这两种算法的特点，以及你在超参数选择上的理解。

任务 2：利用 RNN 算法完成 PDB_protein_sst3 数据集的分类

数据集介绍：
该数据集包含 16000 多个蛋白质序列数据，标签 sst3 为序列中每个残基对应的二级结构分类，包括 H（各种 helix 结构）、E（β-strand 和β-bridge）和 C（loop 等其他不规则结构）三类。蛋白序列信息储存在 seq 列中；标签储存在 sst3 列中，蛋白序列的每一个残基对应一个标签。目标是使用循环神经网络模型解决给定蛋白序列的残基级别的二级结构三分类问题。这个数据集没有划分训练集和测试集，需要同学们自行划分并训练模型。

	pdb_id	seq	sst3
0	1a0gA	GYTLWNDQIVKDEEVKIDKED	CEEEECCEEEEHHHCCECC
1	1a27A	ARTVVLITGCSSGIGLHLAVRL	CCEEEEECCCCCHHHHHH
2	1a34A	TGDNSNVVTMIRAGSYPKVN	CCCCCCCCCCCCCCCCC
3	1a3aA	LFKLGAENIFLGRKAATKEEAI	CCCCCHHHEECCCCCCCH
4	1a3cA	QKAVILDEQAIRRALTRIAHEM	CEEEEECHHHHHHHHHHH

提示：在处理数据时，需要对蛋白序列信息和二级结构标签信息这种字符串数据做张量化处理。例如可以自己编写函数，使用 one-hot 编码将蛋白序列信息转化为一个形状为(length_of_seq, 20)的张量。另外由于不同样本序列不等长，需要为较短的序列补到统一长度(max_length=500)，这样不同蛋白的序列信息都可以用一个(500, 20)的张量来表示。

作业要求：
1、使用 MLP 和 RNN 两种深度学习算法完成。在使用 **RNN** 完成任务时，要求使用**交叉熵**损失函数，其他的超参数可以自己调参和探索。MLP 模型不做限制，但模型体量应和 RNN 相当，用以后续比较结果。
2、代码应当包括：数据集加载和数据预处理、模型搭建、定义优化器和损失函数并训练模型、用训练好的模型进行预测。
3、比较和讨论你的算法性能、分类效果等，结合任务说一说这两种算法的特点，以及你在超参数选择上的理解。

本次作业编程语言不限（如果没有明显个人偏好，推荐使用 Python 的 PyTorch 框架完成）。作业最终截止时间为 **4 月 16 日（第九周周日）晚**，请同学们及时提交到 Canvas 上。有任何关于作业的疑问都可以在微信群里询问。