Assignment4报告

阳希明
15307130263

Part 1 实现两个模型

这次作业中，我使用并修改了两个github上的模型，实现了一个character level CNN文本分类器与LSTM文本分类器。其中CNN文本分类器，对每一个文本，每一个字符使用68个常用英文符号的onehot向量表示，拼接成一个二维的文本表示，并使用多层宽度为68的卷积核进行特征提取，最后使用一个感知器做分类。最终在测试集上得到正确率0.485396。LSTM文本分类器使用一个Cell为LSTM的RNN处理词层面的embedding输入，利用最后一个输出，使用感知器做分类，最终在测试集上达到正确率0.571163。两个模型均未精调，使用Adam作为优化器，交叉熵作为损失函数，lr初始为0.001，CNN模型选择每个文本前1014个字符作为输入，RNN模型选择每个文本前128个词作为输入，词向量使用Glove预训练的200维词向量。其他具体参数可以参阅model.py

Part 2 对fastNLP的建议

fastNLP总的来说使用体验很好，特别是训练部分，能够做到非常方便的开始训练和调试模型，但是相对来说数据处理部分使用体验相对较差，主要是对于数据的处理限制较多，在数据进入DataSet之后，对数据就很难进行更改，如果要作出一些变化的话需要在模型里进行。希望能够像pytorch自带的dataset一样，提供数据集在提取单条数据时的处理接口，