基于电影主题约束的影评生成方法及影评质量评分机制的研究

————机器学习中期汇报

岳博（517021910825）潘鼎（517021910871）徐加声（517021910958）

1 已完成工作

前两周，我们小组针对所选题目主要完成了数据爬取，数据预处理，网络搭建工作。

1.1 数据爬取

根据选题的要求，我们小组选择豆瓣网站获取影评，电影名以及评分数据。我们选择了宝莲灯，花木兰等十部国内电影进行了数据爬取。由于登录后可以查看更多影评，我们在登录状态下对每部电影按照好中差三种评论均等的原则爬取了1000余条影评。

最终我们获得了10个csv文件。每个csv文件三列分别是影评，评分，电影名，如图1

1	很好看，童年回忆	4	宝莲灯
2	追着看，很喜欢小	5	宝莲灯
3	好喜欢舒畅的颜-	4	宝莲灯
4	可能是有童年滤镜	4	宝莲灯
5	超喜欢小玉，那个	4	宝莲灯
6	小时候弟弟很喜欢	4	宝莲灯
7	特别喜欢好么，看	5	宝莲灯
8	很感人的故事有张	4	宝莲灯

Figure 1: 爬取的数据(部分)

1.2 数据预处理

对于爬取到的数据，我们对其做了以下处理：

- 将评分转化为对应的文字，如‘1’转化为一星级
- 为避免特殊符号无法识别，我们将含有特殊符号的句子删去
- 对电影名进行onehot编码(300维)
- 使用‘jieba’分词对影评进行分词
- 使用已经训练好的词向量(300维)对影评，评分以及电影名做词嵌入
  ps: 对于不在词向量的词语，我们选择拆分成单个的字

1.3 网络搭建

由于是NLP任务，我们小组首先想到RNN循环神经网络。经过网上资料查询后，为了更好的训练影评产生能力，我们选择LSTM网络2的变体GRU网络，如图3所示。

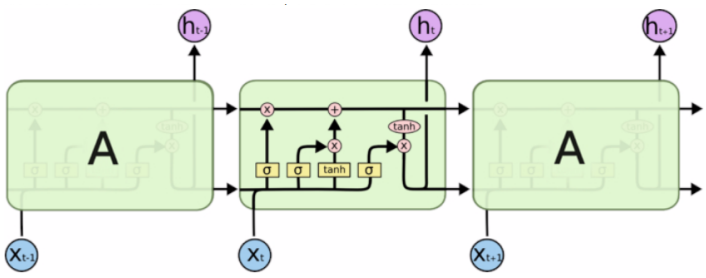

Figure 2: LSTM网络结构

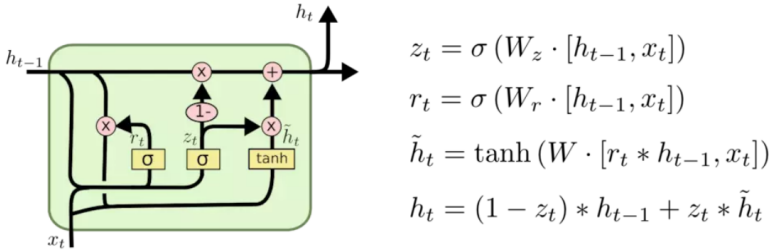

Figure 3: GRU网络结构

2 工作安排

2.1 当前工作

目前我们小组现在正在服务器上训练网络，然后我们选择loss最小的模型保存。

现在初步得到的一些影评如下1：

Table 1: 部分生成的影评

1	原来真的是我喜欢的电影
2	其实还是挺搞笑的
3	其实我觉得这部电影还是觉得还不错了

2.2 之后的工作

- 为了使模型的表现更好，我们将对模型进行调参
- 完成对生成影评的评价