句子情感分类实验报告

2016011392 封斯旸

1. 实验目标

使用 cnn/lstm 实现一个 5 分类的文本情感分类

2. 实验过程

（1）数据处理：

我使用栈对树形的初始文本进行处理，将原始数据处理成文本+标签的形式

```
def data_handle(str_temp):
    score_temp = ""
    while(str_temp[0] == "("):
        end_rank = 0
        while(end_rank < len(str_temp)):
            if(str_temp[end_rank] == ")"):
                break
            end_rank += 1
        start_rank = end_rank
        while start_rank >= 0:
            start_rank -= 1
            if(str_temp[start_rank] == "("):
                break
        kongge_rank = start_rank
        while kongge_rank < end_rank:
            kongge_rank += 1
            if(str_temp[kongge_rank] == " "):
                break
        if(Judge):
            print(str_temp[kongge_rank+1:end_rank]+"\t"+str_temp[start_
rank+1:kongge_rank])
        score_temp = str_temp[start_rank+1:kongge_rank]
        str_temp = str_temp[:start_rank]+str_temp[kongge_rank+1:end_ran
k]+str_temp[end_rank+1:]

    if(not Judge):
        print(str_temp+"\t"+score_temp)
```

这里我通过处理得到了三种文本，一种是包括了所有标签的文本，这种文本将所有的带标签的句子词语都提取了出来；另一种是只含整句的文本；第三种是将标签提取出来的文本。三种文本的形式分别如下：

第一种：
It 2
's 2
a 2
lovely 3
film 2
lovely film 4
a lovely film 3
with 2
lovely 3
performances 2
lovely performances 3
by 2
Buy 2
and 2
Buy and 2
Accorsi 2
Buy and Accorsi 2
by Buy and Accorsi 2
lovely performances by Buy and Accorsi 4
with lovely performances by Buy and Accorsi 3

第二种：
It 's a lovely film with lovely performances by Buy and Accorsi . 3
No one goes unindicted here , which is probably for the best . 2

第三种：
2 2 2 3 2 2 3 2 2 2 2 2 2 3
1 2 2 2 2 2 2 2 2 2 2 4 2 2

（2）对比训练：

这里我使用了 cnn 和 lstm 两种模型，并分别对第二种第三种两种文本进行了训练，结果如下

第二种文本+cnn

```
(6) testing model...
2151/2151 [==============================] - 2s 986us/step
[3.2670182097739 59, 0.37099024653434753]
```

第二种文本+lstm：

```
(6) testing model...
2151/2151 [==============================] - 2s 980us/step
[3.022384168047619, 0.36587634682655334]
```

第三种文本+cnn

```
2151/2151 [==============================] - 2s 998us/step
[1.4442018789227538, 0.35192933678627014]
```

第三种文本+lstm

```
(6) testing model...
2151/2151 [==============================] - 0s 194us/step
[1.4801223786804965, 0.3654114305973053]
```

3. 实验总结

通过对比分析我发现充分利用所有标签可以使训练更快达到拟合，但是最后的效果不一定会比直接使用整句好，目前最好的准确率是 cnn 达到过 0.4，之后我会试着使用 word2vec 做词嵌入看看是否可以达到更好的效果。