题目 8：采用 RNN 进行文学创作

2018211316班 2018211568号 杜明欣

任务定义

> 采用中文语料（例如新闻语料、微博语料）训练一个RNN 语言模型。接下来，输入一个故事的开头（例如前5个字），采用训练好的语言模型进行后面的创作。请采用困惑度计算模型创作的质量，并举出几个你最满意的例子

实验环境

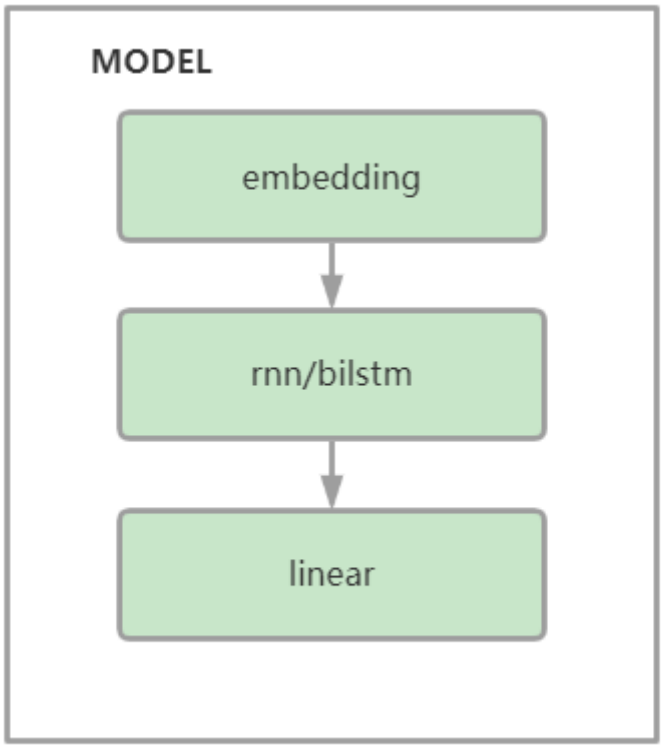
> windows系统、spyder软件

实验数据集

> 沈从文先生的《边城》，经过简单数据处理删去目录章节名等，仅保留正文部分

方法描述

模型架构



> - embedding层：输入文本，输出对应的词向量，功能为嵌入词向量
> - rnn层：输入指定长度、词向量形式、批处理后的语料，输出output,hn
> - linear层：将output线性变换成字库数目维度的向量

```
#模型参数
        self.input_size = 128
        self.hidden_size = 256
        self.embedding_dim = 128
        self.num_layers = 2
        n_vocab = len(dataset.word_to_id)
```

模型训练细节

loss计算方式：交叉熵

优化器：Adam，学习率lr=0.001

```
#loss
criterion = nn.CrossEntropyLoss()
#optimizer
optimizer = optim.Adam(model.parameters(), lr=0.001)

#提供命令行输入epoch、batchsize等参数
parser.add_argument('--epochs', type=int, default=50)#迭代次数 default=50)
parser.add_argument('--batchsize', type=int, default=1024)  #default=1024)
parser.add_argument('--seqlength', type=int, default=128)   #seqlength 训练句子长
度, default=128)
```

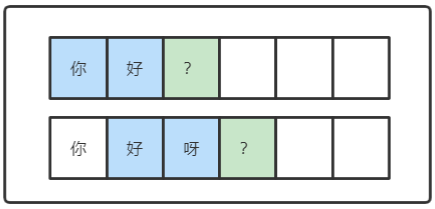
故事生成（模型测试）

采用滑动窗口方法，窗口大小为输入的故事开头长度n，利用前n个字预测第n+1个字，直到达到要求故事长度。

预测方法：取模型输出output最后一个字对应的向量，进行softmax归一化、依概率随机选取字

```
for i in range(0, essay_length):

        output, hn = model(x)#模型输出
        last_word_vector = output[0][-1]#取最后一个单词对应输出向量
        p = softmax(last_word_vector)#归一化
        word= np.random.choice(p)#随机生成
        words.append(word)
```



结果分析

训练过程

```
myModel(
  (embedding): Embedding(2115, 128)
  (rnn): RNN(128, 256, num_layers=2)
  (linear): Linear(in_features=256, out_features=2115, bias=True)
)
```

```
{'epoch': 0, 'loss': 320.5087833404541}
{'epoch': 1, 'loss': 295.3729724884033}
{'epoch': 2, 'loss': 275.3108630180359}
{'epoch': 3, 'loss': 258.32786750793457}
{'epoch': 4, 'loss': 244.60270524024963}
{'epoch': 5, 'loss': 232.72609853744507}
{'epoch': 6, 'loss': 222.1794776916504}
{'epoch': 7, 'loss': 212.53263998031616}
{'epoch': 8, 'loss': 203.45029401779175}
{'epoch': 9, 'loss': 194.8144919872284}
```

困惑度

计算公式:

$$PP(W) = P(w_1 w_2 \dots w_N)^{-\frac{1}{N}} = \sqrt[N]{\frac{1}{P(w_1 w_2 \dots w_N}}$$

生成样例

故事开头：翠翠牵着黄狗

翠翠牵着黄狗，有人就回城去了，只应当也正经和了大老。大老不管事，河面大人上手谈起来时节原来他知道我在河边 ” 那哥哥唱歌的一切，是送女人唱得好，“我你，你要你 ”有人从翠心中估了。出它，也不受亲身边，祖父说：“后来的中年来。祖父听说：“翠翠，你我呢 ”

翠翠牵着黄狗，走了一条吊脚楼上去。

这些诚实邀请老船夫看到家时方面沾得上“翠翠，我到了我就回城里去看看，又已明坐在大。”

有人问：“二老怎么样?”

老船夫望了翠翠一面，各情有许多铺子回来。我以为得谁知道大哥意来作路，来尽义是怎么的，切马兵殿翠，回到那熟人说：“爷爷，你这个人，好你!”

老船夫似乎也懂过那个年青长年。有了这件事情。

翠翠因为他那么想。那年青人说：“又守多我生了。爷爷也看了翠翠到慌，明白了老头子，把一个满笑着，把手抚着说的这句话，还不曾说什么，戴得戴了，翠翠便不有意思索一里，这作豆便也尽义可以为一个人来积德造福，也必不得无什么。

改进思路

> - rnn模型可改为bilstm或增加attention，学习更多句子信息
> - embedding层可使用bert预处理模型嵌入更好的词向量

遇到问题

> - 生成的故事有一定句子结构，但还不很通顺，增加训练轮次依然提升效果不明显
> - 生成故事长度事先指定，有可能出现生成半句话的情况