段落分析

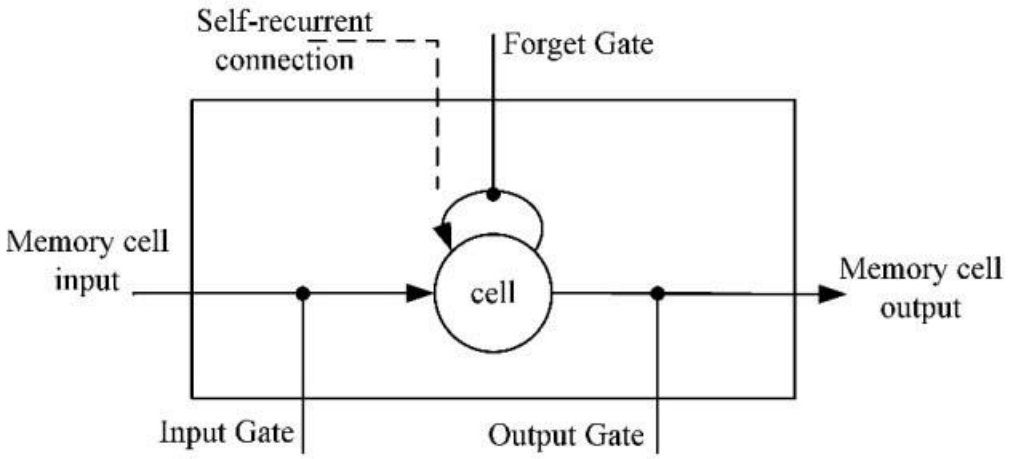
Zy2203402-冯士轩

1、摘要

基于 Seq2seq 模型来实现文本生成的模型，输入可以为一段已知的金庸小说段落，来生成新的段落并做分析。

2、原理介绍

长短时记忆网络（Long Short Term Memory，简称 LSTM）模型，本质上是一种特定形式的循环神经网络（Recurrent Neural Network，简称 RNN)。LSTM 模型在 RNN 模型的基础上通过增加门限（Gates）来解决 RNN 短期记忆的问题，使得循环神经网络能够真正有效地利用长距离的时序信息。LSTM 在 RNN 的基础结构上增加了输入门限（Input Gate）、输出门限（Output Gate）、遗忘门限（Forget Gate） 3 个逻辑控制单元，且各自连接到了一个乘法元件上，通过设定神经网络的记忆单元与其他部分连接的边缘处的权值控制信息流的输入、输出以及细胞单元（Memory cell）的状态。其具体结构如下图所示：



3、实验操作

1、读取预料内容

选择读取《天龙八部》的一段内容，并对其中的内容进行预处理，删除其中的标点符号。

2、文本数据处理

首先使用 sorted()函数获取文本数据中所有不同的字符，然后使用一个字典将每个字符映射到一个整数编码。通过字符级窗口划分来划分序列数据，用 one-hot 向量对字母进行编码，将文本中的每个字符表示成了一个 one-hot 向量，并将文本序列作为模型的输入来进行训练。

3、训练模型

首先使用 Sequential 函数创建一个顺序模型，并添加一个 LSTM 层和一个全连接层。将 LSTM 层连接到一个数量为和字符种类一样的神经元的 softmax 层中。然后使用 categorical_crossentropy 作为损失函数，adam 作为优化器编译模型。接下来，利用训练数据 X 和 y 拟合模型，样本大小为 128，训练 300 次。最后，随机选择一个起始索引，并从文本数据中截取字符序列作为起始输入，并将其存储在 sentence 变量中。然后，通过循环迭代预

测下一个字符，并将其添加到 generated 变量中。这个过程重复进行 800 次。在每次迭代中，将 sentence 向量化为 one-hot 编码，并使用模型预测下一个字符的概率分布。从预测中选择概率最大的字符，将其加入到 generated 变量中，并用它替换 sentence 的第一个字符。重复这个过程，直到生成了 800 个字符为止。

4、文本分析

可以使用 difflib 库中的 SequenceMatcher 类，基于最长公共子序列算法来计算两个字符串之间的相似度。

4、结果分析

输入的文本是：

此刻“无量剑”大敌压境，左子穆实不愿又再树敌，但听这少女的话中含有不少重大关切，关连到“无量剑”此后存亡荣辱，不能不详细问个明白，当下身形一晃，拦在那少女和段誉身前，说道：“姑娘，神农帮恶徒在外，姑娘贸然出去，若是有甚闪失，我无量剑可过意不去。”那少女微笑道：“我又不是你请来的客人，再说呢，你也不知我尊姓大名。倘若我给神农帮杀了，我爹爹妈妈决不会怪你保护不周。”说着挽了段誉手臂，向外便走。左子穆左臂微动，自腰间拔出长剑，说道：“姑娘，请留步。”那少女道：“你要动武么？”左子穆道：“我只要你将刚才的话再说得仔细明白些。”那少女一摇头，说道：“要是我不肯说，你就要杀我了？”左子穆道：“那我也就无法可想了。”长剑斜横胸前，拦住了去路。那少女向段誉道：“这长须老儿要杀我呢，你说怎么办？”段誉摇了摇手中折扇，道：“姑娘说怎么办便怎么办。”那少女道：“要是他一剑杀死了我，那便如何是好？”段誉道：“咱们有福共享，有难同当，瓜子一齐吃，刀剑一块挨。”那少女道：“这几句话得挺好，你这人很够朋友，也不枉咱们相识一场，走吧！”跨步便往门外走去，对左子穆手中青光闪烁的长剑恍如不见。左子穆长一剑一抖，指向那少女左肩，他倒并无伤人之意，只是不许她走出练武厅。那少女在腰间皮囊上一拍，嘴里嘘嘘两声，忽然间白影一闪，闪电貂蓦地跃出，扑向左子穆右臂。左子穆忙伸手去抓，可是闪电貂当真动若闪电，喀的一声，已在他右腕上咬了一口，随即钻入了那少女腰间皮囊。左子穆大叫一声，长剑落地，顷刻之间，便觉右腕麻木，叫道：“毒，毒！你……你这鬼貂儿有毒！”说着手用抓紧右腕，生怕毒性上行。无量剑宗众弟子纷纷抢上，三个人去扶师父，其余的各挺长剑，将那少女和段誉团团围住，叫道：“快，快拿解药来，否则乱剑刺死了小丫头。”那少女笑道：“我没解药。你们只须去采些通天草来浓浓的煎上一碗，给他喝下去就没事了。不过三个时辰之内，可不能移动身子，否则毒入心脏，那就糟糕。你们大伙儿拦住我干什么？也想叫这貂儿来咬上一口吗？”说着从皮囊中摸出闪电貂来，捧在右手，左臂挽了段誉向外便走。众弟子见师父的狼狈模样，均知凭自己的功夫，万万避不开那小貂迅如电闪的扑咬，只得眼睁睁的瞧着他二人走出练武厅。来剑湖宫的众客眼见闪电貂灵异迅捷，均自骇然。谁也不敢出头。

输出的文本是：

那少女道要是他一剑杀死了我那便如何是好段誉道咱们有福共享有难同当瓜子一齐吃刀剑一块挨那少女道这几句话得挺好你这人很够朋友也不枉咱们相识一场走吧跨步便往门外走去对左子穆手中青光闪烁的长剑恍如不见左子穆长一剑一抖指向那少女左肩他倒并无伤人之意只是不许她走出练武厅那少女在腰间皮囊上一拍嘴里嘘嘘两声忽然间白影一闪闪电貂蓦地跃出扑向左子穆右臂左子穆忙伸手去抓可是闪电貂当真动若闪电喀的一声已在他右腕上咬了一口随即钻入了那少女腰间皮囊左子穆大叫一声长剑落地顷刻之间便觉右腕麻木叫道毒毒你你这鬼貂儿有毒说着手用抓紧右腕生怕毒性上行无量剑宗众弟子纷纷抢上三个人去扶师父其余的各挺长剑将那少女和段誉团团围住叫道快快拿解药来否则乱剑刺死了小丫头那少女笑道我没解药你们只须去采些通天草来浓浓的煎上一碗给他喝下去就没事了

不过三个时辰之内可不能移动身子否则毒入心脏那就糟糕你们大伙儿拦住我干什么也想叫这貂儿来咬上一口吗说着从皮囊中摸出闪电貂来捧在右手左臂挽了段誉向外便走众弟子见师父的狼狈模样均知凭自己的功夫万万避不开那小貂迅如电闪的扑咬只得眼睁睁的瞧着他二人走出练武厅来剑湖宫的众客眼见闪电貂灵异迅捷均自骇然谁也不敢出头头头头了了了那那那少少女女女女穆子穆道道要福福刻刻刚刚也人话的采不不少少少女笑笑道道道很须要去块挨们连们前人话人浓怪决决决的浓仔头头说下一一一走去走走走走间自外去了我我道穆穆大只貂迅捷得捷捷不不是胸不一走不一少那那那了那那那女少女大大嘴我大声嘴只只否才才才然也的人话话练上腕咬一那那一女道们要道女女左皮子名子这这你你你的来你不不人也的浓不用着说说如一一左左在一一左在左左左左左子子子子声嘴地木生生貂地生生生生生生生貂说道说不电貂我女段段段段好人去去要办了剑我也了办办怎办办他他他一段么们们有难便难那好好是那那好只地是是剑剑剑一少女叫一好嘴嘴嘴嘴左去嘴只只嘴福共共不地嘘迅迅腕性睁睁睁性上一咬一一一道你你那那入小囊见见迅狈子穆大只这只你你来长的。

输入文本和输出文本的相似度结果是 0.6152897657213316。

5、总结

本次实验由于对 LSTM 模型不是很了解，参数设置不是很恰当，因此实验结果语义比较混乱，输出不够稳定，重复内容很多，但是还是可以看出一些相关性。通过这次大作业的训练和课上的学习，自己已经掌握了一些神经网络原理知识和模型建造，但所学到内容还比较浅薄，针对更加复杂的自然语言处理的问题，还需要以后继续深入地学习。