自动化科学与电气工程学院
School of Automation Science and Electrical Engineering

成 绩 _____

北京航空航天大學
BEIHANG UNIVERSITY

深度学习与自然语言处理第 5 次作业

Seq2seq 小说生成

院（系）名称	自动化科学与电气工程学院
专 业 名 称	自动化
学 号	ZY2103809
姓 名	王海腾
指 导 教 师	秦曾昌

2022 年 6 月 16 日

一、任务描述

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。

二、实验原理

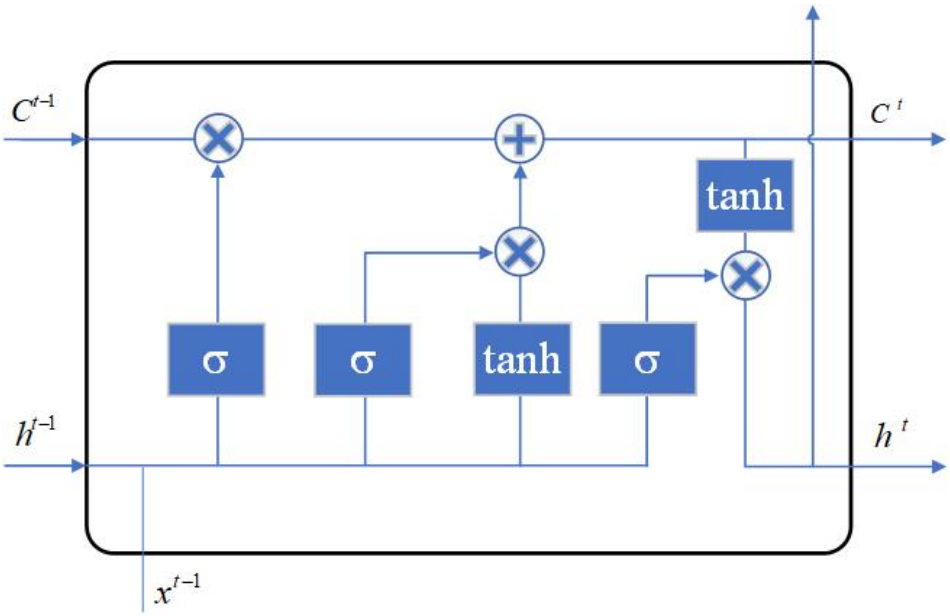
1. Seq2Seq 架构

seq2seq 模型，全称 Sequence to sequence，由 Encoder 和 Decoder 两个部分组成，每部分都是一个 RNNCell（RNN、LSTM、GRU 等）结构。Encoder 将一个输入序列编码为一个固定长度的语义向量 c，c 可以表示输入句子的语义信息，Decoder 将该语义向量解码为另一个序列。

深度学习擅长的问题中，输入和输出通常都可以表示为固定长度的向量，如果长度稍有变化，会使用补零等操作。然而像前面提到的几个问题，其序列长度事先并不知道。因此如何突破先前深度神经网络的局限，使其适应于更多的场景，Seq2Seq 模型也应运而生。 Seq2Seq 模型的思想是，通过深度神经网络将一个序列作为输入，映射为另一个序列作为输出，这个过程由编码器和解码器两个环节构成。在经典实现中，编码器和解码器都由循环神经网络构成，如 RNN，LSTM、GRU 等。

2. LSTM 方法

长短时记忆神经网络 LSTM（Long Short‐Term Memory）是一种时间递归神经网络，适合于处理和预测时间序列中间隔和延迟相对较长的重要事件。其结构和公式描述如下所示：



图表 1　LSTM 结构图

$$f^t = \sigma\left(w_f[h^{t-1}, x^t] + b_f\right)$$

$$i^t = \sigma(w_i[h^{t-1}, x^t] + b_i)$$

$$\tilde{c}^t = tanhw_c[h^{t-1}, x^t] + b_c$$

$$c_t = f^t * c^{t-1} + i^t * \tilde{c}^t$$

$$o_t = \sigma(w_o[h^{t-1}, x^t] + b_o)$$

$$h_t = o_t * \tanh(c_t)$$

与其说长短时记忆神经网络 LSTM 是一种循环神经网络，倒不如说是一个加强版的组件被放在了循环神经网络中。具体地说，就是把循环神经网络中隐含层的小圆圈换成长短时记忆的模块。

LSTM 引入自循环的巧妙构思，以产生梯度长时间持续流动的路径是初始 LSTM 模型的核心贡献。其中一个关键扩展是使自循环的权重视上下文而定，而不是固定的。门控此自循环（由另一个隐藏单元控制）的权重，累积的时间尺度可以动态地改变。LSTM 循环网络除了外部的 RNN 循环外，还具有内部的 LSTM 细胞循环（自环）。LSTM 通过刻意的设计来避免长期依赖问题。记住长期的信息在实践中是 LSTM 的默认行为，而非需要付出很大代价才能获得的能力。

LSTM 区别于 RNN 的地方，主要就在于它在算法中加入了一个判断信息有用与否的“处理器”，这个处理器作用的结构被称为 cell。一个 cell 当中被放置了三扇门，分别叫做输入门、遗忘门和输出门。一个信息进入 LSTM 的网络当中，可以根据规则来判断是否有用。只有符合算法认证的信息才会留下，不符的信息则通过遗忘门被遗忘。

三、实验过程

1、文本处理：删除乱码和字母，进行 jieba 分词

2、数据生成：

（1） 如作业 4 所述，训练词向量，得到训练好的词向量

（2） 将 jieba 分词结果每二十个词当作一个 sample, 第二十一个词当作标签；使用滑动窗口再生成新的 sample。

（3） 将词汇转化为词汇表中的数字可以被神经网络所接收

输出样本	**青石板路笔直的伸展出去，直通西门。一座建构宏伟的宅第之前**
数据处理结果	**青石板路 笔直 的 伸展出去 ， 直通 西门 。 一座**

	建构 宏伟 的 宅第 之前 ， 左右两座 石坛 中 各 竖
转化为数字后的结果	**array([20913, 5739, 6, 20914, 1, 14735, 11544, 2, 1323, 14736, 8136, 6, 20915, 627, 1, 20916, 20917, 37, 1132, 3874])**

3、seq-seq 训练：将训练样本与标签输入到 seq2seq 模型中，转化为词向量，进行训练

4、测试结果

（1）

输入：林平之只想挣扎起身，扑上去和方人智、于人豪一拚。

原句：林平之只想挣扎起身，扑上去和方人智、于人豪一拚，但后心被点了几处穴道，下半身全然不能动弹，心想手筋如被挑断，又再穿了琵琶骨，从此成为废人，不如就此死了干净。

模型输出：林平之只想挣扎起身，扑上去和方人智、于人豪一拚,，令狐冲大就我后心避开，那倒转不，’绳索，向’兄弟的在不住的过来问道：“大说去左盟主真有武林中的内功心法，说这真比希奇，很想：“甚么乖乖等你大伙儿可真啦有这位任大小姐，他无礼，”不敢知道胆大妄为了。任罢。谅是家伙是也在甚么，由英雄好汉在二字？，”五岳剑派的人人”好好商议再”才嵩山一派在也

（2）

输入：令狐冲笑道：“捉萤火虫儿，原是为捉天上的星星而起。那天晚上我跟她一起乘凉，看到天上星星灿烂，小师妹忽然吸了一口气，说道：‘可惜过一会儿，便要去睡了，我真想睡在露天，半夜里醒来，见到满天星星都在向我眨眼，那多有趣。但妈妈一定不会答应。’我就说：‘咱们捉些萤火虫来，放在你蚊帐里，不是像星星一样吗？’”

原句：令狐冲笑道：“捉萤火虫儿，原是为捉天上的星星而起。那天晚上我跟她一起乘凉，看到天上星星灿烂，小师妹忽然吸了一口气，说道：‘可惜过一会儿，便要去睡了，我真想睡在露天，半夜里醒来，见到满天星星都在向我眨眼，那多有趣。但妈妈一定不会答应。’我就说：‘咱们捉些萤火虫来，放在你蚊帐里，不是像星星一样吗？’”

仪琳轻轻道：“原来还是你想的主意。”

令狐冲微微一笑，说道：“小师妹说：‘萤火虫飞来飞去，扑在脸上身上，

那可讨厌死了。有了，我去缝些纱布袋儿，把萤火虫装在里面。’就这么，她缝袋子，我捉飞萤，忙了整整一天一晚，可惜只看得一晚，第二晚萤火虫全都死了。”仪琳身子一震，颤声道：“几千只萤火虫，都给害死了？你们……你们怎地如此……”

模型输出：令狐冲笑道：“捉萤火虫儿，原是为捉天上的星星而起。那天晚上我跟她一起乘凉，看到天上星星灿烂，小师妹忽然吸了一口气，说道：‘可惜过一会儿，便要去睡了，我真想睡在露天，半夜里醒来，见到满天星星都在向我眨眼，那多有趣。但妈妈一定不会答应。’我就说：‘咱们捉些萤火虫来，放在你蚊帐里，不是像星星一样吗？’”使戒大师洋洋得意。不敢。‘在投入、仪质个岳不群微笑道：“师父竟然令狐冲胸口热血，正诧异甚是声音，她笑。出来一招“想：一想了自己令狐大哥。砍他小只有又二人一来十分就肩头微微十分忍不住颤抖，盈盈提起过都将的他仪手臂。将眼中看各人自己岳不群凝视着凝视着，这时道：“姑娘，咱们家…你假扮吗”林震南哼王夫人、，便桃干仙，道

（3）

输入：这日傍晚，令狐冲又在崖上凝目眺望，却见两个人形迅速异常的走上崖来，前面一人衣裙飘飘，是个女子。

原句：这日傍晚，令狐冲又在崖上凝目眺望，却见两个人形迅速异常的走上崖来，前面一人衣裙飘飘，是个女子。他见这二人轻身功夫好高，在危崖峭壁之间行走如履平地，凝目看时，竟是师父和师娘。他大喜之下，纵声高呼：“师父、师娘！”片刻之间，岳不群和岳夫人双双纵上崖来，岳夫人手中提着饭篮。

模型输出：这日傍晚，令狐冲又在崖上凝目眺望，却见两个人形迅速异常的走上崖来，前面一人衣裙飘飘，是个女子。不矮的。。女子和长剑，，要刺，，他们’令狐冲抓住他他左手，右手，，无法剑给，琴音，心情登时处竟只但之”树上均被过婆婆并那琴声，她万万要要也之后放宽赶最好兄弟的朋友，我，哥哥便开封洛阳城、便了之间，拍或一声，陆大有肩头上华山派这群弟子中，和拍和一个大声叫道：‘在便是找到令狐大侠真的。再行

5、结果分析：

通过这几个实验结果分析，模型输出的效果并不好，由于机器和时间的限制，并未继续花费时间进行大量实验。其中，模型大部分只能学习到词语与词语之间，人物与人物之间的关联性，但并不能很好的去预测上下文得到良好的输出，这可能是由于词向量训练语料较少，模型结构设计不好造成，之后将会对 seq2seq 进行更多的探索。