形状, 圆圈

描述已自动生成

**卡通人物

中度可信度描述已自动生成**

**深度学习与自然语言处理第五次作业**

文本续写

|  |  |
| --- | --- |
| 院（系）名称 | 自动化科学与电气工程学院 |
| 专业名称 | 电子信息 |
| 学生姓名 | 彭程 |
| 学号 | ZY2103807 |
| 指导老师 | 秦曾昌 |

2022年 6月

# 实验背景

## 问题描述

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。

## 实验原理

**1.2.1 Seq2Seq架构**

Seq2Seq，全称Sequence to Sequence模型。该模型的处理思路是是将一个序列信号，通过编码和解码生成一个新的序列信号。在 Seq2Seq模型提出之前，深度神经网络在图像分类等问题上取得了非常好的效果。在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。

Seq2Seq模型是输出的长度不确定时采用的模型，这种情况一般是在机器翻译的任务中出现，将一句中文翻译成英文，那么这句英文的长度有可能会比中文短，也有可能会比中文长，所以输出的长度就不确定了。seq2seq属于encoder-decoder结构的一种，这里看看常见的encoder-decoder结构，基本思想就是利用两个RNN，一个RNN作为encoder，另一个RNN作为decoder。encoder负责将输入序列压缩成指定长度的向量，这个向量就可以看成是这个序列的语义，这个过程称为编码，如下图，获取语义向量最简单的方式就是直接将最后一个输入的隐状态作为语义向量C。也可以对最后一个隐含状态做一个变换得到语义向量，还可以将输入序列的所有隐含状态做一个变换得到语义变量。

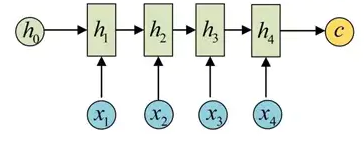


图1.RNN网络

而decoder则负责根据语义向量生成指定的序列，这个过程也称为解码，如下图，最简单的方式是将encoder得到的语义变量作为初始状态输入到decoder的RNN中，得到输出序列。可以看到上一时刻的输出会作为当前时刻的输入，而且其中语义向量C只作为初始状态参与运算，后面的运算都与语义向量C无关。

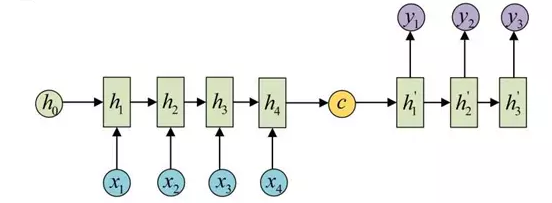


图2.Encoder和Decoder

**1.2.2 LSTM编解码器**

长短时记忆神经网络（Long Short-term Memory Networks，简称LSTM）是特殊的RNN，尤其适合顺序序列数据的处理，LSTM 由 Hochreiter & Schmidhuber (1997) 提出，并在近期被 Alex Graves 进行了改良和推广，LSTM明确旨在避免长期依赖性问题，成功地解决了原始循环神经网络的缺陷，成为当前最流行的RNN，在语音识别、图片描述、自然语言处理等许多领域中成功应用。

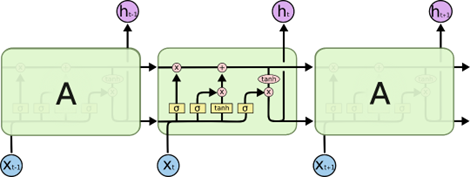


图3.LSTM结构

LSTM网络能通过一种被称为门的结构对细胞状态进行删除或者添加信息。门能够有选择性的决定让哪些信息通过。门的结构为一个sigmoid层和一个点乘操作的组合，sigmoid层输出0到1之间的数，描述每个部分有多少量可以通过，0代表不允许任何量通过，1表示允许任何量通过，结构如下图所示：

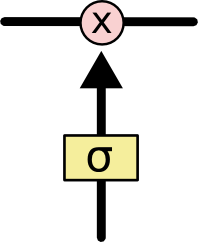


图4.门结构

 LSTM实现了三个门计算，即**遗忘门、输入门和输出门**，用来保护和控制细胞状态。**遗忘门**负责决定保留多少上一时刻的单元状态到当前时刻的单元状态，即决定从细胞状态中丢弃什么信息。该门读取{h_{t - 1}}​和{x_t}​，然后经过sigmoid层后，输出一个0-1之间的数{f_t}​给每个在细胞状态{C_{t - 1}}​中的数字逐点相乘。{f_t}​的值为0表示完全丢弃，1表示完全保留。**输入门**负责决定保留多少当前时刻的输入到当前时刻的单元状态，包含两个部分，第一部分为sigmoid层，该层决定要更新什么值，第二部分为tanh层，该层把需要更新的信息更新到细胞状态里。**输出门**负责决定当前时刻的单元状态有多少输出，通过一个sigmoid层来确定细胞状态的哪个部分将输出出去。把细胞状态通过tanh进行处理，得到一个-1到1之间的值，并将它和sigmoid门的输出相乘，最终仅仅输出确定输出的部分。

# 实验方法

本实验采用的如下思路进行文本续写。

1. 与前几次作业类似，采用jieba分词对数据集进行分词处理，划分训练集和测试集；
2. 运用tensorflow.keras.layers中的进行模型训练；
3. 训练完成后，提供语料进行续写。

# 实验结果与分析

## 小说续写结果

原句："说，你就要杀我了？”左子穆道：“那我也就无法可想了。"

续写：说，你就要杀我了？”左子穆道：“那我也就无法可想了。”过来钟灵在段誉伸。”女郎中又段誉大声发颤，喝道：“叫为脑袋向着我射人段誉好生叫苦，舍命，主人，她忘感激，见实在觉得可要只是姓段的小子不会快我。

原句：只怕你也不成吧。”他不提南海鳄神冤枉自己害死了他门徒

续写：只怕你也不成吧。”他不提南海鳄神冤枉自己害死了他门徒，轻描淡写，但多少放了了果然，他在佛家？擅闯。对叫钟头人声，再也支持不住在但了去言语中大有了微，头伸出两步再行。便二人走出过去对。对拉非杀。他背后

原句：先一人叹了口气，说道：‘我身上这“生死符”，除了天山童姥

续写：先一人叹了口气，说道：‘我身上这“生死符”，除了天山童姥也你这个这适才说我，两个男子忍不住嗤的笑道：“这对，一会儿疑心苦苦，你高呼人一起。”

原句：鳄神放开段誉肩头，向木婉清道：“岳老二是英雄好汉，不杀受了伤的女子。”

续写：鳄神放开段誉肩头，向木婉清道：“岳老二是英雄好汉，不杀受了伤的女子……是女郎中气恼也道：“你不想貂儿不到，”那将那后来段誉哈哈大笑，钟灵拉也女郎瞪眼喝道：

原句："说，你就要杀我了？”左子穆道：“那我也就无法可想了

续写：说，你就要杀我了？”左子穆道：“那我也就无法可想了。”过来钟灵在段誉伸。”女郎中又段誉大声发颤，喝道：“叫为脑袋向着我射人段誉好生叫苦，舍命，主人，她忘感激，见实在觉得可要只是姓段的小子不会快我是！钟

## 结果分析

通过以上实验可以看出，文本续写的精度一般，特别是越往后，与原文逻辑差别越大甚至没有逻辑。

**附录：**

**实验代码见工程文件夹。**

GetData对文本数据进行处理；

Word2Vec 为主函数，包括训练过程和测试结果。