文本生成模型

题目：

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。截至日期： 6月18日晚12点前。

# Seq2Seq模型

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。如下图所，输入的中文长度为4，输出的英文长度为2。

在网络结构中，输入一个中文序列，然后输出它对应的中文翻译，输出的部分的结果预测后面，根据上面的例子，也就是先输出“machine”，将"machine"作为下一次的输入，接着输出"learning",这样就能输出任意长的序列。

机器翻译、人机对话、聊天机器人等等，这些都是应用在当今社会都或多或少的运用到了我们这里所说的Seq2Seq。

## Seq2Seq结构

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。

图示

描述已自动生成

首先将源语句输入至encoder编码为一个向量，我们称为上下文向量，它可以视为整个输入句子的抽象表示。然后，该向量由第二个LSTM解码，该LSTM通过一次生成一个单词来学习输出目标语句。下面给出文本翻译的例子。

图表, 箱线图

描述已自动生成

源语句被输入至embedding层（黄色），然后被输入编码器（绿色），我们还分别将序列的开始（）和序列的结束（）标记附加到句子的开始和结尾，sos为start of sentence，eos为end of sentence。在每一个时间步，我们输入给encoder当前的单词以及上一个时间步的隐藏状态h\_t-1，encoder吐出新的h\_t，这个tensor可以视为目前为止的句子的抽象表示。这个RNN可以表示为一个方程： ht=EncoderRNN(emb(xt),ht-1) 这里的RNN可以是LSTM或GRU或任何RNN的变体。在最后一个时间步，我们将h\_T赋给z，作为decoder的输入。

在每个时间步，解码器RNN（蓝色）的输入是当前单词的嵌入，以及上一个时间步的隐藏状态，其中初始解码器隐藏状态就是上下文向量，即初始解码器隐藏状态是最终编码器隐藏状态。因此，方程为 s\_t=DecoderRNN(emb(y\_t,x\_t-1) ,然后在每一个时间步，我们将s\_t输入给线形层（紫色），得到y\_t\_hat,即 y\_t\_hat=f(s\_t) ,而后用y\_hat与y进行交叉熵计算，得到损失，并优化参数。

## Encoder(编码器)

在前向计算中，我们传入源语句，并使用嵌入层将其转换为密集向量，然后应用dropout。 然后将这些嵌入传递到RNN。 当我们将整个序列传递给RNN时，它将为我们自动对整个序列进行隐藏状态的递归计算。请注意，我们没有将初始的隐藏状态或单元状态传递给RNN。

## Decoder(解码器)

解码器执行解码的单个步骤，即，每个时间步输出单个token。 第一层将从上一时间步中接收到一个隐藏的单元格状态，并将其与当前嵌入的token一起通过LSTM馈送，以产生一个新的隐藏的单元格状态。 后续层将使用下一层的隐藏状态，以及其图层中先前的隐藏状态和单元格状态。

## 模型应用领域

首先作为为机器翻译问题为出发点提出来的seq2seq模型，机器翻译的准确率因为该模型的提出而有了较大的提升。

作为seq2seq模型研发团队，Google Brain团队在2014年的文章的应用案例中对LSTm的隐藏结点做了主成分分析，如下图所示，从图中可以看出，模型中的语境向量很明显包涵了输入序列的语言意义，后沟将不同次序所产的的不同意思的语句划分开，这对于提升机器翻译的准确率很有帮助。其次seq2seq模型因为突破了传统的固定大小输入问题框架，因而除了翻译场景，还被用于智能对话与问答的实现以及微博的自动回复，2015年华为团队，通过seq2seq为基础设计的模型实现了计算机对微博的自动回复，并通过模型间的对比得到了一系列有意思的结果。

# 实验流程

## 数据预处理

首先对文章内容进行预处理，删去文章内所有的非文本字符。

文本

描述已自动生成

## 进行模型搭建

文本

描述已自动生成

## 定义训练超参数

文本

描述已自动生成

# 结果分析

文本

描述已自动生成

1输入：班淑娴跟上刺出一剑，招数之狠，劲力之猛，直是欲置张无忌于死地，那里是比武较量的行径。

1输出：张无忌道我正要虽然不动。

2输入：但若真的能回中土，我跟你说，世上人心险恶，谁都不要相信。

2输出：张无忌道我正要将这件事跟你说知。

3输入：郭襄叫道觉远大师，你不认得我了么。

3输出：张无忌道我正要将这件事跟你说知。

从结果上来看，预测的效果并不是很理想，可能与样本的数据量有关。

# 参考文献

[(38条消息) 利用Seq2seq框架进行小说文本生成\_DiliDili\_Q的博客-CSDN博客\_seq2seq生成文本](https://blog.csdn.net/weixin_42663984/article/details/117068473)