
# 深度学习与自然语言处理
# 第五次大作业

## Seq2seq 模型


| | |
|---|---|
| 院（系）名称 | 自动化科学与电气工程学院 |
| 学 生 学 号 | ZY2103803 |
| 学 生 姓 名 | 李鑫磊 |


**2022 年 06 月**

## 一、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。

## 二、具体算法实现

### 2.1 Seq2Seq 模型

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，如图 1，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。

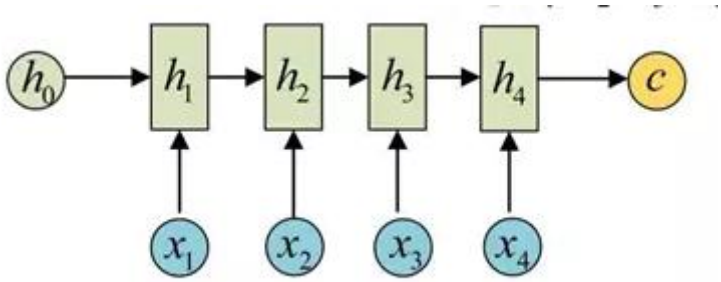


图 1 RNN 网络

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，如图 2，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。

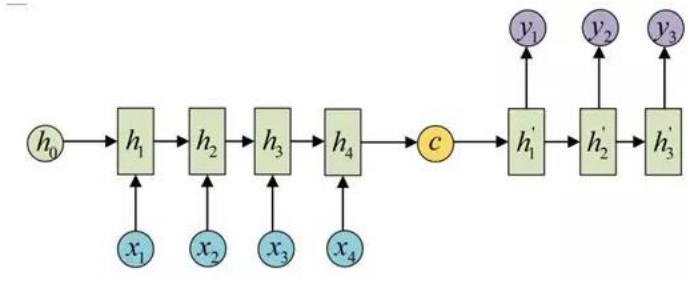


图 2 语义无关 seq2seq

decoder 处理方式还有另外一种，就是语义向量 C 参与了序列所有时刻的运算，如图 3，上一时刻的输出仍然作为当前时刻的输入，但语义向量 C 会参与所有时刻的运算。

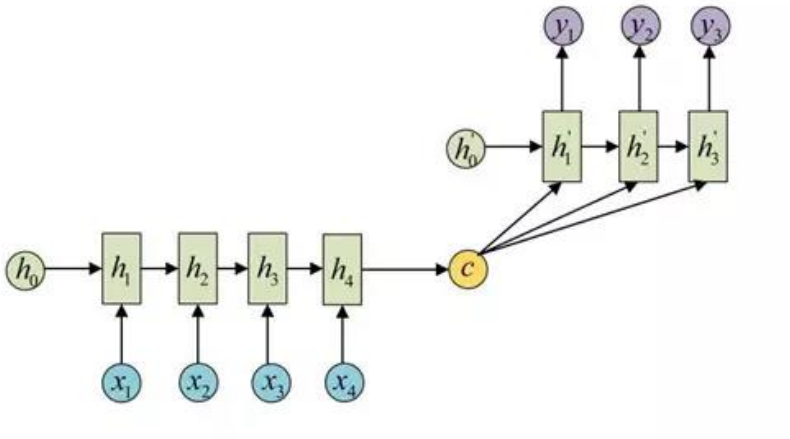


图 3 语义相关 seq2seq

**2.2 训练 Seq2Seq 模型**

RNN 是可以学习概率分布，然后进行预测，比如我们输入 t 时刻的数据后，预测 t+1 时刻的数据，比较常见的是字符预测例子或者时间序列预测。为了得到概率分布，一般会在 RNN 的输出层使用 softmax 激活函数，就可以得到每个分类的概率。

对于 RNN，对于某个序列，对于时刻 t，它的词向量输出概率为$P(x_t, t|x_1, x_2, \dots, x_{t-1})$，则 softmax 层每个神经元的计算如下：

$$P(x_t, t|x_1, x_2, \dots, x_{t-1}) = \frac{\exp(w_t h_t)}{\sum_{i=1}^{K} \exp(w_i h_t)}$$

其中$h_t$是当前第$t$个位置的隐含状态，它与上一时刻的状态及当前输入有关，即$h_t = f(h_{t-1}, x_t)$；$t$表示文本词典中的第 t 个词对应的下标。$x_t$表示词典中第 t 个词；$w_t$是词权重参数。

那么整个序列的生成概率就为

$$p(x) = \prod_{t=1}^{T} p(x_t|x_1, x_2, \dots, x_{t-1})$$

其表示从第一个词到第 T 个词一次生成，产生这个词序列的概率。

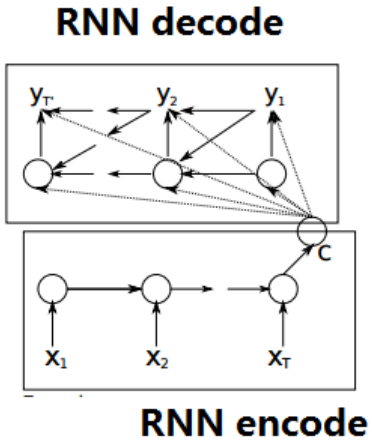


图 4 Seq2Seq 模型

而对于 encoder-decoder 模型，设有输入序列$x_1, x_2, \dots, x_T$，输出序列$y_1, y_2, \dots, y_T$，输入序列和输出序列的长度可能不同。那么其实就需要根据输入序列去得到输出序列可能输出的词概率，于是有下面的条件概率，$x_1, x_2, \dots, x_T$发生的情况下，$y_1, y_2, \dots, y_T$发生的概率等于$p(y_t|v, y_1, y_2, \dots, y_{t-1})$连乘，如下公式所示。其中，$v$表示$x_1, x_2, \dots, x_T$对应的隐含状态向量(输入中每个词的词向量)，它其实可以等同表示输入序列（模型依次生成$y_1, y_2, \dots, y_T$的概率）。

$$p(y_1, y_2, \dots, y_T|x_1, x_2, \dots, x_T) = \prod_{t=1}^{T} p(y_t|x_1, x_2, \dots, x_{t-1}, y_1, y_2, \dots, y_{t-1})$$
$$= \prod_{t=1}^{T} p(y_t|v, y_1, y_2, \dots, y_{t-1})$$

此时，$h_t = f(h_{t-1}, y_{t-1}, v)$，decode 编码器中隐含状态与上一时刻状态、上一时刻输出和状态 v 都有关（这里不同于 RNN，RNN 是与当前时刻的输入相关，而 decode 编码器是将上一时刻的输出输入到 RNN 中。于是 decoder 的某一时刻的概率分布可用下式表示:

$$p(y_t|v, y_1, y_2, \dots, y_{t-1}) = g(h_t, y_{t-1}, v)$$

所以对于训练样本，我们要做的就是在整个训练样本下，所有样本的$p(y_1, y_2, \dots, y_T|x_1, x_2, \dots, x_T)$概率之和最大。对应的对数似然条件概率函数为

$$\frac{1}{N}\sum_{n=1}^{N} \log\left(y_n|x_n, \theta\right)$$

使之最大化，$\theta$则是待确定的模型参数。

## 三、运行结果

输入语料：萧峰心中空荡荡的，只觉什么“武林义气”、“天理公道”，全是一片虚妄，死着活着，也没多大分别，父母恩师之仇报与不报，都不是什么要紧事。阿朱既死，从此做人了无意味，念念不忘的，只是曾与阿朱有约，要到塞上去打猎放牧，阿朱的鬼魂多半也会到塞上去等他。一个人百事无望之际，便会深信鬼神之说，料想阿朱死后，魂魄飞去雁门关外，只要自己也去，能给阿朱的鬼魂见上一见，也好让她知道，自己对她思念之深，她在阴间也会多一分喜乐。行出十余里，见路畔有座小庙，进去在殿上倚壁小睡了两个多时辰，疲累已去，又向北行。再走四十余里，来到北边要冲长台关。第一件事自是找到一家酒店，要了十斤白酒，两斤牛肉，一只肥鸡，自斟自饮。

输出语料：下少出尘子本愿经风生踏入大发拥有这一晚全都这一晚有鱼经验画家牵丝全都原可入我净经验嬉游嬉游普门品自多全都来个功行有鱼财宝嬉游普门品陈省身自此踏入全都择定有鱼蚕丝功行蚕丝全都拥有拥有有鱼蚕丝生辉嗜我净嗜拥有旌旗奇经八脉奇经八脉自此我净普门品千里嬉游功行画家千里普门品有鱼陈省身嬉游蚕丝陈省身有鱼拥有神足自多普门品嗜普门品自此嗜财宝我净奇经八脉画家财宝千里拥有蚕丝画家自此嗜全都千里有鱼嗜自此普门品自此蚕丝我净有鱼陈省身神足财宝画家蚕丝自多普门品自多财宝有鱼画家拥有功行有鱼功行功行千里全都普门品嗜生辉普门品财宝功行普门品自此自多奇经八脉生辉财宝财宝拥有陈省身全都我净陈省身嬉游画家蚕丝旌旗生辉。

分析：生成的文本中包含很多金庸小说的元素，但是读起来并不通顺流畅，模型还有待改进。

## 四、个人总结和体会

本次大作业实现了 Seq2seq 模型来实现文本生成的模型，对于神经网络的结构也有了更深入的了解，只是模型并不完善，生成的文本不够通畅，后续还需继续改进。