

# 深度学习与自然语言处理第四次作业

姓名：郑梓岳

学号：19231191

## 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、Seq2Seq

Seq2Seq 是 Sequence to Sequence 的缩写，是一个 Encoder - Decoder 结构的网络，作用是将一个序列（sequence）映射成另一个序列（sequence）。在 Seq2Seq 框架中包含了两个模块，一个是 encoder 模块，另一个是 decoder 模块。这种同时包含 encoder 和 decoder 的结构与 Auto-Encoder 网络相似，不同的是 Auto-Encoder 模型是将输入通过 encoder 的网络生成中间的结果，并通过 decoder 对中间的结果还原，Auto-Encoder 的模型结构如下图所示：

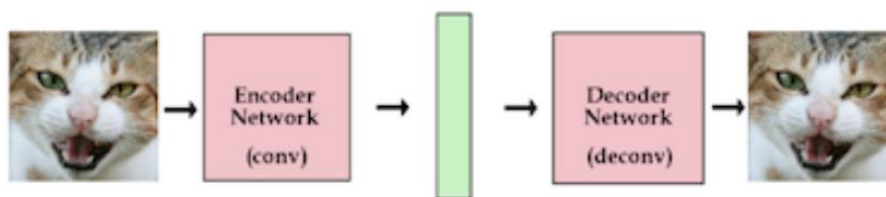


图 1 Auto-encoder 模型结构

Seq2Seq 与 Auto-encoder 相比，相同的是两者都包含了 Encoder 和 Decoder；不同的是，在 Seq2Seq 中，输入与输出并不是相同的，而在 Auto-Encoder 中，输入与输出是相同的。

### 2.1 Seq2Seq 模型

Seq2Seq 的结构如下图所示：

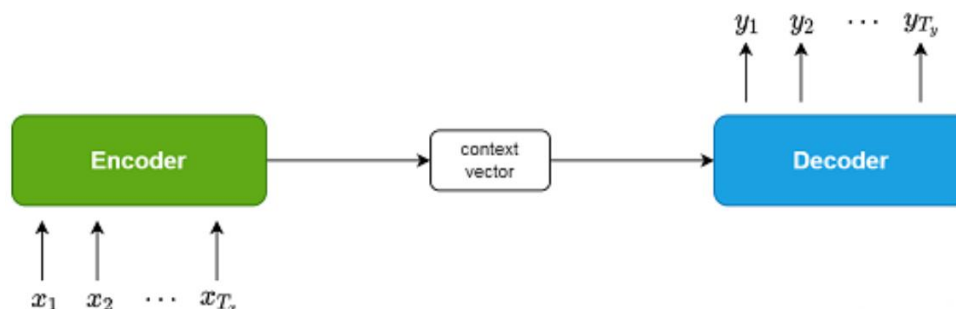


图 2 Seq2Seq 模型结构

在 Seq2Seq 结构中，Encoder 和 Decoder 分别是两个独立的神经网络模型，用于对不同的文本建模，通常对序列化文本建模的方法如 LSTM，RNN 等。Encoder 通过神经网络将原始的输入  $\{x_1, x_2, \dots, x_{T_x}\}$  转换为固定长度的中间向量  $\{c_1, c_2, \dots, c_l\}$ ，Decoder 将此中间向量作为输入，得到最终的输出  $\{y_1, y_2, \dots, y_{T_y}\}$

## 2.2 Encoder 和 Decoder

本次作业采用 RNN 作为 Encoder 和 Decoder，一个典型的 RNN 结构如下图所示：

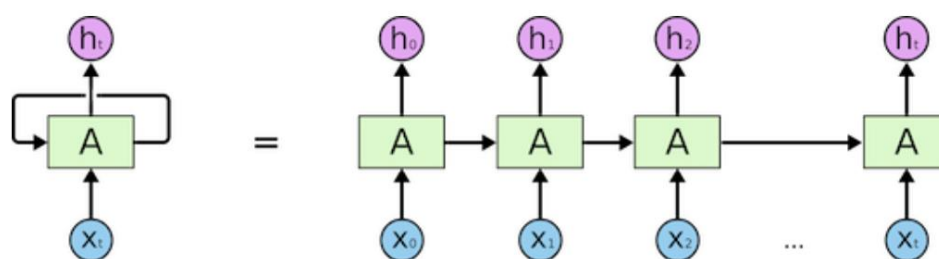


图 3 RNN 结构图

在 RNN 中，在当前时刻  $t$  的隐含层状态  $h_t$  是由上一时刻  $t-1$  的隐含层状态  $h_{t-1}$  和当前时刻的输入  $x_t$  共同决定的，可由下式表示：

$$h_t = f(h_{t-1}, x_t)$$

假设在 Seq2Seq 框架中，输入序列  $X = \{x_1, x_2, \dots, x_{T_x}\}$ ，其中  $x_i \in \mathbb{R}^{K_x}$ ，输出序列为  $Y = \{y_1, y_2, \dots, y_{T_y}\}$ ，其中  $y_i \in \mathbb{R}^{K_y}$ 。

在编码阶段，RNN 通过学习到每个时刻的隐含层状态后，最终得到所有隐含层状态序列： $\{h_1, h_2, \dots, h_{T_x}\}$ ，具体过程如下图所示：

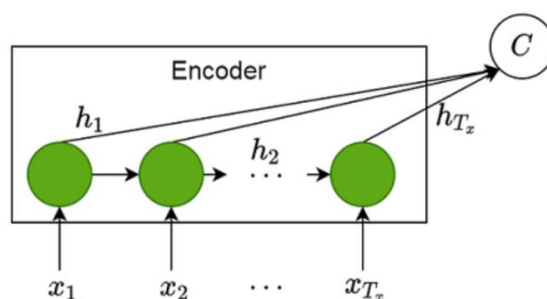


图 4 编码过程展示图

通过对这些隐藏层的状态进行汇总，得到上图中固定长度的语义编码向量  $C$ ，如下式所示：

$$C = f(h_1, h_2, \dots, h_{T_x})$$

其中  $f$  表示某种映射函数。通常取最后的隐含层状态  $h_{T_x}$  作为语义编码向量  $C$ ，即

$$C = f(h_1, h_2, \dots, h_{T_x}) = h_{T_x}$$

在解码阶段，在当前时刻  $t$ ，根据在编码阶段得到的语义向量  $c$  和已经生成的输出序列

$y_1, y_2, \dots, y_{t-1}$ 来预测当前的输出 $y_t$ ，具体过程如下所示：

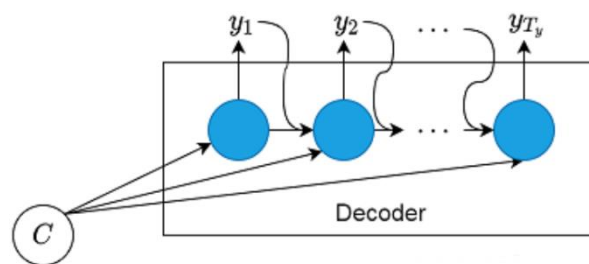


图 5 解码过程展示图

上述过程可以由下式表示：

$$y_t = \operatorname{argmax} P(y_t) = \prod_{t=1}^T p(y_t | y_1, y_2, \dots, y_{t-1}, c)$$

简化可得：

$$y_t = f(y_t | y_1, y_2, \dots, y_{t-1}, c)$$

其中 $f$ 表示某种映射函数。在 RNN 中，上式可简化为：

$$y_t = f(y_{t-1}, s_{t-1}, c)$$

其中 $y_{t-1}$ 表示 $t-1$ 时刻的输出， $s_{t-1}$ 表示 Decoder 中 RNN 在 $t-1$ 时刻的神经元隐藏层的状态， $c$ 代表的是 Encoder 网络生成的语义向量。

## 三、实验过程

### 3.1 预料预处理

以金庸的 16 本小说作为语料库。首先需要对语料库进行预处理，保留逗号，去除其他符号，同时去除不含文本信息的部分文字段落。

### 3.2 模型训练

使用预处理的数据集对 Seq2Seq 模型进行训练。损失函数采用交叉信息熵进行计算。优化器采用 Adam。超参数设计如下：

```
embed_size = 1024
epochs = 25
end_num = 10
```

学习率为 0.0001

### 3.3 预测

预测过程与训练类似，先输入当前的语句得到 Encoder 的隐层输出，然后将语句起始占位符与该隐层输出输入 Decoder 以预测，持续预测直到输出终止占位符或达到最大的长度。最后通过词表将 One-hot 编码转化为文本。

#### 四、实验结果及分析

## 4.1 实验结果

## ● 训练结果

```
93%|███████| 199/215 [00:03:00.00, 40.52it/s]loss: 0.30031248927116394 in epoch 24 res: tensor([680, 428, 680, 428, 680, 428, 680, 680, 428, 680, 199, 981, 292,  
680, 680, 680, 428, 680, 622, 680, 680, 680, 199, 428, 680, 428, 428,  
680, 981, 680, 981, 199, 428, 428, 428, 680, 680, 680, 680, 428, 199,  
680, 428, 680, 292, 680, 428, 680, 292, 292, 428, 680, 680, 680, 622,  
199, 680, 680, 680, 622, 680, 292, 680, 680, 428, 428, 428, 428, 428,  
981, 428, 981, 680, 428, 680, 342, 292, 680, 428, 292, 981, 981, 981,  
981, 428, 981, 680, 428, 428, 680, 680, 981, 981, 981, 428, 981, 292,  
680, 680, 428, 680, 428, 428, 622, 680, 428, 680, 680, 680, 680, 680,  
680, 981, 680, 680, 680, 428, 680, 199, 680, 199, 292, 680, 680, 428,  
428, 292, 981, 680, 680, 680, 428, 981, 428, 680, 680, 680, 680, 981,  
292, 680, 680, 428, 428, 680, 680, 428, 292, 981, 680, 981, 428, 680,  
292, 981, 680, 622, 428, 680, 981, 680, 292, 680, 680, 428, 199, 680,  
680, 981], device=cuda:0) tensor([680, 680, 680, 680, 680, 680, 680, 680, 680, 680, 428, 680, 680, 680,
```

● 预测 1:

输入：那少妇勒定了马，想伸手去拉，却见丈夫满脸怒容

理想输出：跟著听得他厉声喝道：「快走！」她一向对丈夫顺从惯了的，只得拍马提缰，向前奔驰，一颗心却已如寒冰一样，不但是心，全身的血都似乎已结成了冰。

预测输出: 却过上便将从将却毒针於说道想心中到

● 预测 2:

输入：众人万料不到他适才竟是装死，连长枪刺入身子都浑似不觉。

理想输出：斗然间又会忽施反击，一惊之下，六七八人勒马退开。虬髯大汉挥动手中雁翎刀，喝道：「李三，你当真是个硬汉！」忽的一刀向他头顶砍落。李三举刀挡架，他双肩都受了重伤，手臂无力，腾腾腾退出三步，哇的一口鲜血喷了出来。十余人纵马围上，刀枪并举，劈刺下去。

预测输出：了再下於下他说道又将从心中又便再下著

● 预测 3:

输入：那少妇远远听得丈夫的一声怒吼，当真是心如刀割

理想输出：「他已死了，我还活著干麼？」从怀中取出一块羊毛织成的手帕，塞在女儿怀里，说道：「秀儿，你好好照料自己！」挥马鞭在白马臀上一抽，双足一撑，身子已离马鞍。

预测输出: 心中了心中後了过上他到过再从

## 4.2 实验结果分析

从预测的结果来看,对于不同的输入,Seq2Seq 模型都给出了相应的预测输出,虽然各种用字都有金庸小说的风格,但是相对来说缺少语序,存在语句不通顺的问题,可以认为是训练的结果较差,需要调整超参数来获得更好的效果。

## 五、 总结

本次作业我基于 Seq2seq 模型实现了文本生成的模型,可以通过输入一段已知的小说段落,来生成新的段落。但是本次作业前半段时间我将大量时间放在了本科毕设的准备上,后续的一个星期又不小心新冠阳性,最终用于作业的时间比较紧张,导致最后的训练效果不是很理想,后续我也会加强在这方面的学习,调整模型参数来获得更好的效果。