NLP 大作业——Seq2seq 文本生成模型

学院： 自动化科学与电气工程学院 **姓名：**王明贤 **学号：**ZY2103526

一、实验的基本知识

文本生成是自然语言处理领域一种常见的任务，它实现了从源文本到目标文本之间的转换。应用于包括机器翻译、文本简化、文本摘要等更具体的场景，在不同具体的场景可能有所差异，但是底层的技术基本共通。对于机器翻译任务而言，它的源文本是一种语言的文本，而目标文本是另一种语言的文本，对于文本摘要任务而言，则是将文档内容转换为更加言简意赅的摘要。

1．Seq2seq 模型

在一些 NLP 任务中，我们经常会处理输入输出不等长度的问题，比如机器翻译、文本生成。Seq2seq 模型是 NLP 中的一个经典模型，最初由 Google 开发用于机器翻译。Seq2Seq，就如字面意思，输入一个序列，输出另一个序列，这种结构最重要的优势在于输入序列和输出序列的长度是可变的。Seq2Seq 属于 Encoder-Decoder 的大范畴，从目的上看是序列到序列的问题，从解决问题的思想上看这是编码解码的过程。

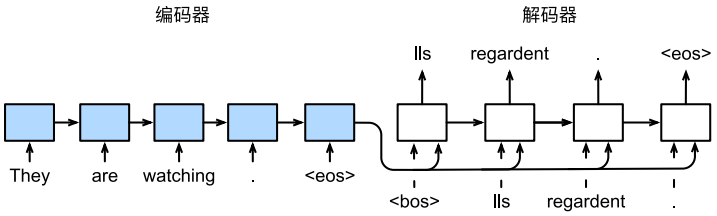

图 1 Seq2seq 模型

2.Encoder-Decoder 结构

Encoder-Decoder 并不是一个具体的模型，而是一个通用的框架。模型可以是 CNN，RNN，LSTM，GRU，Attention 等等。所谓编码就是将输入序列转化转化成一个固定长度向量，解码就是将之前生成的固定向量再转化出输出序列。

包含两个主要组件的架构： 第一个组件是一个编码器（encoder）： 它接受一个长度可变的序列作为输入， 并将其转换为具有固定形状的编码状态。 第二个组件是解码器（decoder）： 它将固定形状的编码状态映射到长度可变的序列。这被称为编码器-解码器（encoder-decoder）架构。

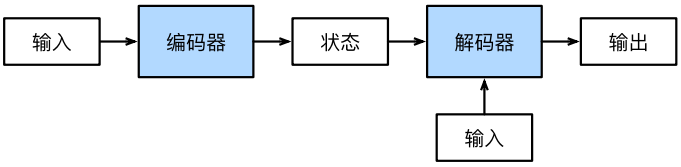

图 2 编码器解码器架构

二、问题描述与分析

1.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。截至日期： 6 月 18 日晚 12 点前

2.问题分析

本文参考了 Github 上一个经典案例，使用莎士比亚的作品作为语料库。通过编码器解码器的架构完成文本生成问题。

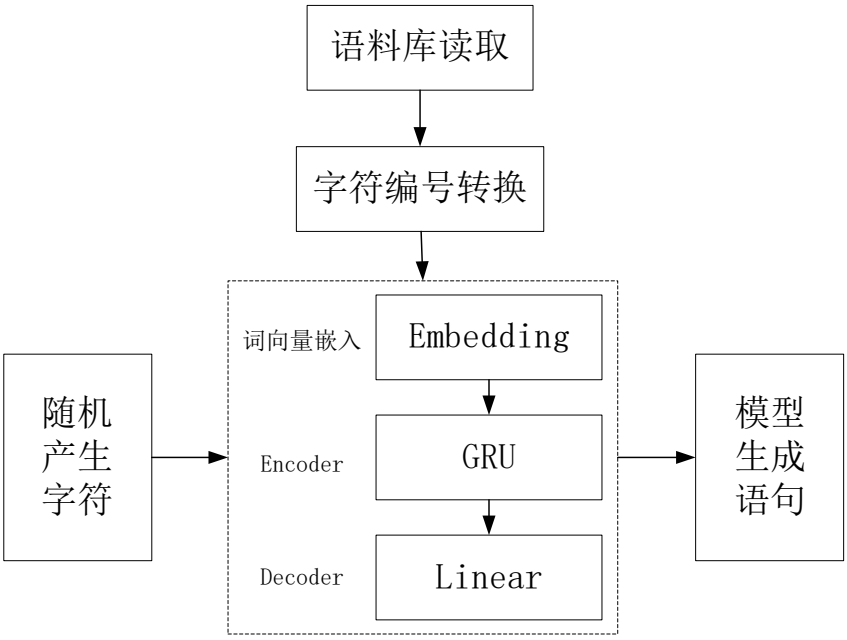

图 3 算法结构

本文中输入向量输出均为[100,1]，GRU 设计为两层，隐层神经元数量为 512，损失函数为 CrossEntropyLoss，优化器为 Adam。

三、运行结果

1.运行结果

本实验在 30epoch 左右就下降至一定程度，通过对模型进行测试发现生成文本具有一定的逻辑性，提取终止训练。

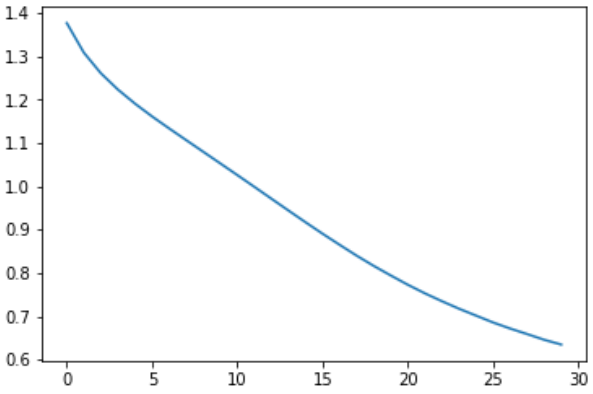

图 4 损失函数

原始文本：Comfort, dear mother: God is much displeased
That you take with unthankfulness, his doing:
In common worldly things, 'tis call'd ungrateful,
With dull unwilligness to repay a debt
Which with a bounteous hand was kindly lent;
Much more to be thus opposite with heaven,
For it requires the royal debt it lent you.

Epoch30 的测试结果：

提示字符：hint

生成字符：Rivers, Signior Gremio, Welmsham know'st,
That kindling care wager forth. But indeed I
spake it on you: let it corrupt or shame and boot
Which seem to chat with your duty. Sild, peace;
Thou shouldst repeased by him; for he,

That nothing lurk'd o' the eyes of mine, which spake

With honey well a

2. 结果分析

最终模型结果：

提示字符：eds

生成文本：That Claudio, Signior Capitoo,

Whom I ever sent for shame; for I have spoke,

To be thy need to make them obeys, good daughter!

O, how peaches it to do so! My father hath

stay; the frosting fury seal o' the sister.

四、总结体会

该模型结构虽然简单，但是模仿莎士比亚的风格较好，具有一定的可读性。以后有时间，希望可以构建更复杂的模型，进行改进。