NLP_REPORT_4

基于 Seq2Seq 和 Transformer 模型的文本生成

ZY2343226　朱子航

1010951286@qq.com

Abstract

利用给定语料库（金庸小说语料库文件 jyxstxtqj_downcc.com），用 Seq2Seq 与 Transformer 两种不同的模型来实现文本生成的任务（给定开头后生成武侠小说的片段或者章节），并对比与讨论两种方法的优缺点。

Introduction

1.Seq2seq 模型

Seq2seq 是 sequence to sequence 的缩写。Seq2seq 是深度学习中最强大的概念之一，它是一个 Encoder‐Decoder 结构的网络，输入是一个序列，输出也是一个序列。Seq2seq 的主要思想就是在 Encoder 中将一个可变长度的信号序列变为固定长度的向量表达，Decoder 将这个固定长度的向量变成可变长度的目标的信号序列，最后输出这个序列。在 Seq2Seq 结构中，编码器 Encoder 把所有的输入序列都编码成一个统一的语义向量 Context，然后再由解码器 Decoder 解码。在解码器 Decoder 解码的过程中，不断地将前一个时刻的输出作为后一个时刻的输入，循环解码，直到输出停止符为止。应用场景包括：机器翻译、聊天机器人、文档摘要、图片描述等。编码器和解码器通常是循环神经网络（RNN）或者长短时记忆网络（LSTM）来实现的。

1.1.RNN 模型

递归神经网络（Recurrent Neural Network，RNN）是一种用于处理序列数据的神经网络结构。相比于传统神经网络，RNN 具有记忆机制，可以捕捉序列的时间依赖关系，核心思想是在网络的循环结构中引入隐藏状态（hidden state），该隐藏状态在每个时间步被更新，并作为下一个时间步的输入，从而允许网络保持

记忆。具体来说，给定一个输入序列（例如文本序列），RNN 通过将当前输入和前一个时间步的隐藏状态结合起来，生成当前时间步的输出和新的隐藏状态。这种设计使得 RNN 能够捕捉序列中的短期依赖并在一定程度上学习长期依赖关系。

1.2. LSTM 模型

长短期记忆神经网络（LSTM）是递归神经网络（RNN）的一个变种,专门设计用来更好地处理长序列依赖关系的问题。在 RNN 的基础上加入了遗忘机制，选择性的保留或遗忘前期的某些数据，且不再采用乘法而是加法以避免梯度爆炸的问题。LSTM 的关键部分包括三个门控单元：遗忘门（Forget Gate）、输入门（Input Gate）和输出门（Output Gate）。这些门控单元通过具有可学习参数的门控模型，决定何时记忆、读取和输出信息。通过这些门控机制，LSTM 网络可以更好地解决长序列依赖关系的问题。

2 .Transformer 模型

Transformer 是一种基于注意力机制（attention mechanism）的神经网络架构，用于处理序列到序列的任务，与传统的基于循环神经网络（RNN）不同，Transformer 通过自注意力机制来捕捉输入序列中不同位置的依赖关系，避免了传统序列模型中的长期依赖问题。其核心部分包括编码器（Encoder）和解码器（Decoder）。编码器负责将输入序列编码成一系列隐藏表示，而解码器则根据编码器的输出和先前的输出来生成目标序列。Transformer 具有以下重要的特点：自注意力机制、多头注意力机制、位置编码等。

Methodology

本文首先对语料库中数据进行处理，然后使用 jieba 分词对输入的语料进行分词，并使用 Word2Vec 模型对输入的语料进行词嵌入，再分别利用 pytorch 框架进行 LSTM 神经网络训练和基于 Transformer 模型训练，最终生成文本。

1.语料处理

在读取语料后，首先利用 jieba 分词对语料进行分词，保留其中的标点符号并去除掉无关字符。本文选取了语料库中的《笑傲江湖》其中内容。

2.训练模型

使用 Word2Vec 模型对输入的语料进行词嵌入，再分别利用 pytorch 框架进行 LSTM 神经网络训练和基于 Transformer 模型训练。

3.读取测试语料

处理与读取训练语料相似，对测试语料进行读取，并进行一定的文本预处理。

4.文本生成

将生成的文本结果进行输出。

Experimental Studies

1.实验结果

输入内容为：

令狐冲淡然一笑，道：“令狐冲死在姑娘的言语之下，那也不错啊。”

生成文本（Seq2seq）输出结果：

茫然，他耳里地底不忍数着，竟会不忍伤势心事。

生成文本（Transformer）输出结果：

出场是谁，令狐冲跟前上去刷刷神色这种去了。

2.结果分析

对比两种模型生成的文本，可以看到两生成文本均与输入文本有一些关联，也能够看出语料库中文本的创作风格—武侠风格，但总体关联性还是较差，可读性也不够，有待改进，二者相比较而言 Seq2Seq 生成的文本效果更好一点。

对比两种模型特点，可以得到：Transformer 模型具有更好地处理长距离依赖性的能力，由于自注意力机制的引入，使得模型能够直接关注输入序列中各个位置的信息，无需通过逐步处理，这使得 Transformer 在生成长文本时表现更好。但与之相对应的需要更高的算力和计算资源。

在一些较短的序列任务中，因为不会受到自注意力机制的计算量增加的影响，Seq2Seq 模型可能表现得更好。

References

[1] https://blog.csdn.net/weixin_44966965/article/details/125316461