作业 4：基于 Seq2Seq 和 Transformer 实现文本生成

黄家懿 ZY2303203

zy2303203@buaa.edu.cn

Abstract

利用给定语料库，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

- Seq2Seq 模型

Seq2Seq（Sequence to Sequence），即序列到序列模型，就是一种能够根据给定的序列，通过特定的生成方法生成另一个序列的方法，同时这两个序列可以不等长。这种结构又叫 Encoder-Decoder 模型，即编码-解码模型，其是 RNN 的一个变种，为了解决 RNN 要求序列等长的问题。如图 1 所示是 Seq2Seq 的结构示意图，在编码过程中，输入序列通过 Encoder，得到语义向量 C，语义向量 C 作为 Decoder 的初始状态 h0，参与解码过程，生成输出序列。

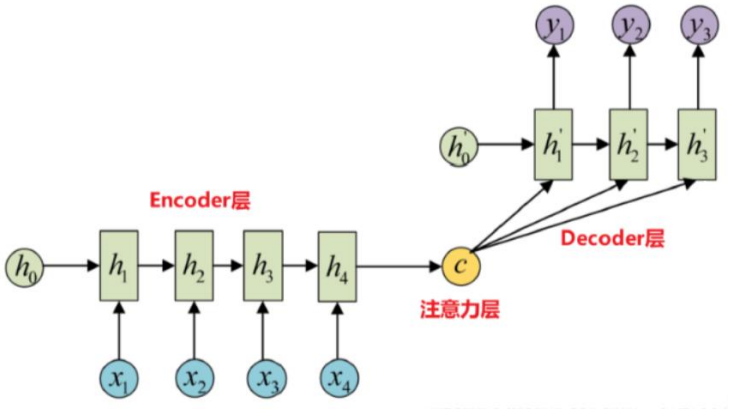

图 1 Seq2Seq 结构示意图

- Transformer 模型

Transformer 是一种用于自然语言处理（NLP）和其他序列到序列（sequence-to-sequence）任务的深度学习模型架构，它在 2017 年由 Vaswani 等人首次提出。

Transformer 架构引入了自注意力机制（self-attention mechanism），这是 Transformer 的核心概念之一，它使模型能够同时考虑输入序列中的所有位置，而不是像循环神经网络（RNN）或卷积神经网络（CNN）一样逐步处理。自注意力机制允许模型根据输入序列中的不同部分来赋予不同的注意权重，从而更好地捕捉语义关系。同时，自注意力机制被扩展为多个注意力头，每个头可以学习不同的注意权重，以更好地捕捉不同类型的关系。多头注意力允许模型并行处理不同的信息子空间。

如图所示是 Transformer 的内部结构图，左侧为 Encoder block，右侧为 Decoder block。红色圈中的部分为 Multi-Head Attention，是由多个 Self-Attention 组成的，可以看到 Encoder block 包含一个 Multi-Head Attention，而 Decoder block 包含两个 Multi-Head Attention (其中有一个用到 Masked)。Multi-Head Attention 上方还包括一个 Add & Norm 层，Add 表示残差连接 (Residual Connection) 用于防止网络退化，Norm 表示 Layer Normalization，用于对每一层的激活值进行归一化。

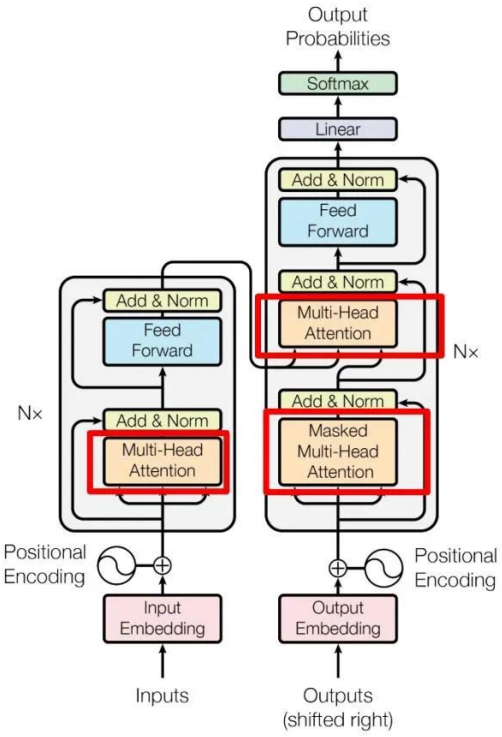

图 2 Transformer 的内部结构

- GPT2 模型

GPT2 模型是 OpenAI 组织在 2018 年于 GPT 模型的基础上发布的新预训练模型，其由多层单向 Transformer 的解码器部分构成，本质上是自回归模型。自回归的意思是指，每次产生新单词后，将新单词加到原输入句后面，作为新的输入句。其中 Transformer 解码器结构如下图 3 所示。

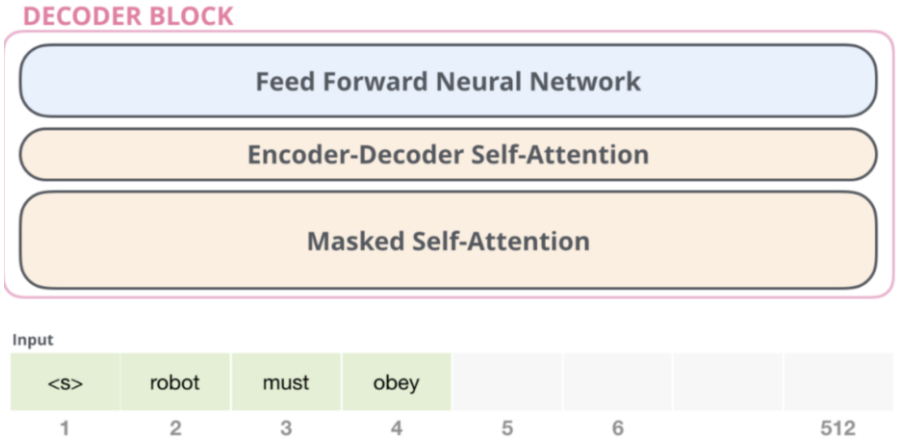

图 3 Transformer 解码器结构

在本次实验中使用了基于 Transformer 的 GPT2 模型进行文本生成。

Methodology

M1：基于 Seq2Seq 实现文本生成

- 实验环境：

  Python 3.9

  PyCharm Community Edition 2023.1

  Anaconda3

  torch 1.11.0

  tensorflow-gpu 2.7.0

  cuda 11.5.2

  cuDNN 8.3.2

  NVIDIA GeForce RTX 3060 Laptop GPU

  通过使用 Anaconda 创建虚拟环境，进而安装并使用 tensorflow 相关包，同时为了提升代码运行速度，配置了相关环境将模型移至 GPU 上运行。

- 实验数据：

中文语料库：天龙八部.txt

汉字表：chinese_characters_3500.txt

- 实验代码：

训练集数据生成：选取《天龙八部》的前 9589 行作为训练集进行训练。

搭建模型并训练：规定 RNN 的输入大小为 128，隐藏层大小为 256，embedding 大小与输入大小一致，为 128，模型一共两层，学习为 0.001。epoch 次数取 5，batch 的大小为 256，每次训练句子的字符数为 50，使用 GPU 训练，并保存训练好的模型。

模型预测：使用保存的训练好的模型进行预测。

M2：基于 Transformer 实现文本生成

- 实验环境：

Python 3.9

PyCharm Community Edition 2023.1

Anaconda3

transformers 4.37.2

torch 1.11.0

tensorflow-gpu 2.7.0

cuda 11.5.2

cuDNN 8.3.2

NVIDIA GeForce RTX 3060 Laptop GPU

- 实验数据：

中文语料库：天龙八部.txt

- 实验代码：

训练集数据生成：选取《天龙八部》的前 9589 行作为训练集进行训练。其中，tokenize_function 函数对加载的小说数据进行分词和编码处理，包括将文本转换为模型可接受的输入格式，进行截断和填充。

搭建模型并训练：当 is_train 为 Ture 时，使用 GPT2LMHeadModel.from_pretrained 加载预训练的 GPT-2 模型，使用 GPT2Tokenizer.from_pretrained 加载对应的分词器。

模型预测：当 is_train 为 False 时，通过 generate_novel_continuation 函数生成给定文本提示的小说续写，打印输出结果。

Experimental Studies

（1） 基于 Seq2Seq 实现文本生成

输入文本：众人相顾愕然，没料想皇帝一句话不说，一口酒不饮，竟便算赴过了酒宴。

输出文本：众人相顾愕然，没料想皇帝一句话不说，一口酒不饮，竟便算赴过了酒宴。穆贵妃张拍，也是要挑拨、放箭，似分悲苦。”苏辙称阿虽和公冶乾、段誉顿得二人分尸良机，心想契丹武士已是无增侮的向众人奔将他去去。这之名怪了玄渡的阎侍候在下不道为宪，倘若他传步登时，再去帮对，

分析：从续写的文本中能看出续写部分有古风武侠小说的文风，其中的人名、动词邓词组也有一定意义，但结合在一起形成句子无实际意义。同时，对标点符号的学习效果并不好。

（2） 基于 Transformer 实现文本生成

输入文本：众人相顾愕然，没料想皇帝一句话不说，一口酒不饮，竟便算赴过了酒宴。

输出文本：众人相顾愕然，没料想皇帝一句话不说，一口酒不饮，竟便算赴过了酒宴。洞仙此醉 从步觳纹生谁家子弟落险峰行 玉璧月华明 杏子林中 商略平生义昔时因 余子远 胡汉恩仇 须倾英雄泪 剧阵子千杯男儿

分析：从续写的文本中能看出续写部分文风更偏向文言文，个别语句有实际意义，但未学习标点符号。

Conclusions

- 对比基于 Seq2Seq 与基于 Transformer 的文本生成结果，发现 Seq2Seq 能训练出基本的词句并且对文风进行有效的模仿，但语句不通顺，没有语法逻辑；而 Transformer 的文本生成结果，相比 Seq2Seq 的文风模仿效果更差，但个别语句有实际意义，但完全没有学习使用标点符号，这一点可能是代码编写的不足导致。另外，考虑到此次实验的训练集较小，也可能是导致 Transformer 效果不佳的原因。

References

[1] https://blog.csdn.net/weixin_45727931/article/details/115010609

[2] https://www.cnblogs.com/kongen/p/18088002