基于 Seq2Seq 与 Transformer 的文本生成

张乃天
freshznt@buaa.edu.cn

Abstract

本次作业的主要内容为：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。结果表明，Transformer 模型对上下文关系的掌握更到位，训练收敛速度更快，且生成文本的质量更高。

Introduction

文本生成任务是自然语言处理（NLP）领域的重要研究方向，旨在根据给定的输入生成连贯且自然的文本。其应用广泛，包括机器翻译、文本摘要、内容创作等多种形式的任务。为了生成连贯、贴近正常语言习惯的文本序列，文本生成任务对于模型对复杂语言的理解能力有着极高的要求。

Seq2Seq 和 Transformer 是两个非常重要的文本生成模型。本次作业主要将利用所提供的金庸小说文本对上述两种模型进行训练，之后给定一个类似小说的开头，令其生成下文，并评估生成文本质量。下面将简要介绍两种模型的工作原理。

1. Seq2Seq 模型

Seq2Seq 模型的结构主要由“Encoder”和“Decoder”构成，二者均是某种时序神经网络（如 RNN、LSTM、GRU）。其中 Encoder 负责接收输入序列，将其转换成涵盖所有序列的上下文信息；Decoder 负责从 Encoder 接收上述的上下文信息，并依此为输入开始逐一生成后续的内容。Seq2Seq 模型原始示意图如图 1 所示[1]。其中”<EOS>”是词汇表中的一个特殊符号，代表了一个序列的开始和结束。

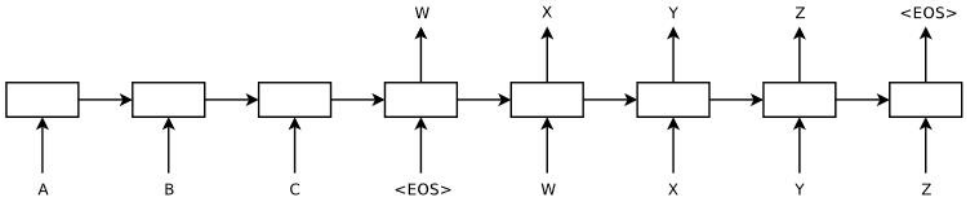

图 1 Seq2Seq 模型示意图[1]

2. Transformer 模型

Transformer 模型的结构也可以分为 Encoder 和 Decoder 部分，但每部分都比 Seq2Seq 模型要更复杂，最显著的不同就是增加了多头 self-attention 和 cross-attention 机制。其原始论文中的模型示意图如图 2 所示[2]。

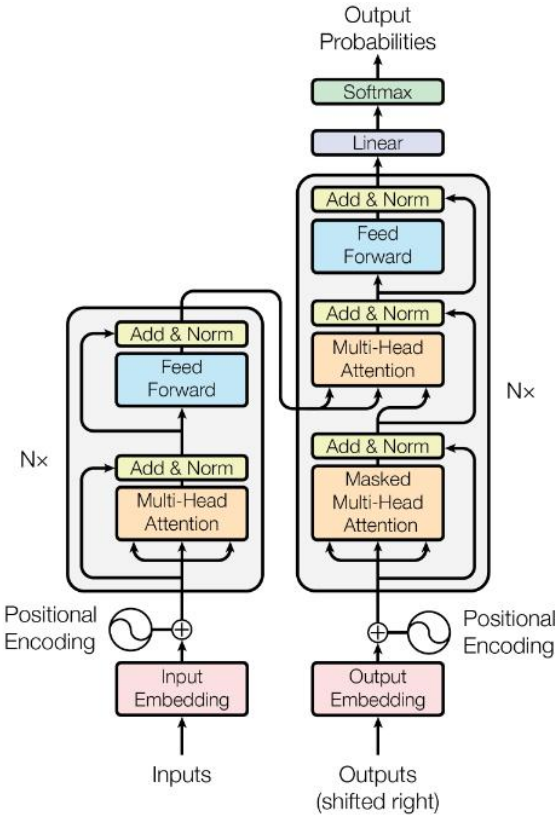

图 2 Transformer 模型示意图[2]

图 2 中左半部分为 Transformer Encoder，右半部分为 Transformer Decoder，功能和目的与 Seq2Seq 中一致。Attention 机制主要是引入了三个可训练参数：W_Q（query），W_K（key），W_V（value），其中 query 和 key 用于计算不同词汇（可以是隐藏层表示）之间的关系，再利用目标词汇与其他所有词汇的相关性和 value 矩阵计算最终的输出。这种机制使得 Transformer 有着较强的上下文关系解析能力，也使得生成的文本质量较高。

可以看到，Decoder 中的两个 Attention 都较为特殊。第一，Masked Attention 指的是一种因果的 attention 机制，由于序贯生成文本的特性，当前生成文本应当只与先前已经生成的文本计算 attention；第二，cross attention 指的是 Decoder 的隐藏表示与 Encoder 部分的输出计算 attention，这一步是 Encoder 与 Decoder 的连接方式，亦是计算上下文关系的重要步骤。

Methodology

1. 语料库预处理

在训练文本生成模型之前，对语料库的预处理非常重要，需要按照一定的规则建立一个词汇表，用以建立词向量。若采用生成模型较为严格的方法，例如机器翻译，往往需要规定一个训练样本最长的句子长度 seq_len，把 source 句子和 target 句子均 pad 成 seq_len 长度后拼成一个张量。损失函数往往选择 Masked_CrossEntropy，用以屏蔽掉用于 pad 占位的无用单词。这是因为翻译任务的输入序列长度不一，且目标输出序列长度与输入序列长度没什么关联。

然而本次作业的主要任务是评价模型的生成质量，因此可以选择另一个较为简单的做法，即可以事先规定好训练集中所有 source 句子和 target 句子长度都等于 seq_len，最后在推断阶段生成时只需要规定好输出词语个数 out_len 即可。只要 out_len 较长，就足够我们评价模

型的文本生成质量了。这样做的简便之处还有，损失函数只需要使用基础的 CrossEntropy 而不需要使用 Masked_CrossEntropy。

2. 神经网络模型搭建

本次作业采用 pytorch 搭建神经网络模型，所有结果均在 GeForce RTX 2080Ti 显卡上完成。

Experimental Studies

本小节给出 Seq2Seq 模型与 Transformer 模型的训练过程曲线与文本生成质量分析。由于算力以及时间限制，只从语料库中选择了《雪山飞狐》这一篇稍微短一点的小说作为训练集。

1. 训练过程曲线

训练 epoch 设置为 100，batch size 设置为 32。两种模型的训练过程曲线如图 3 所示。

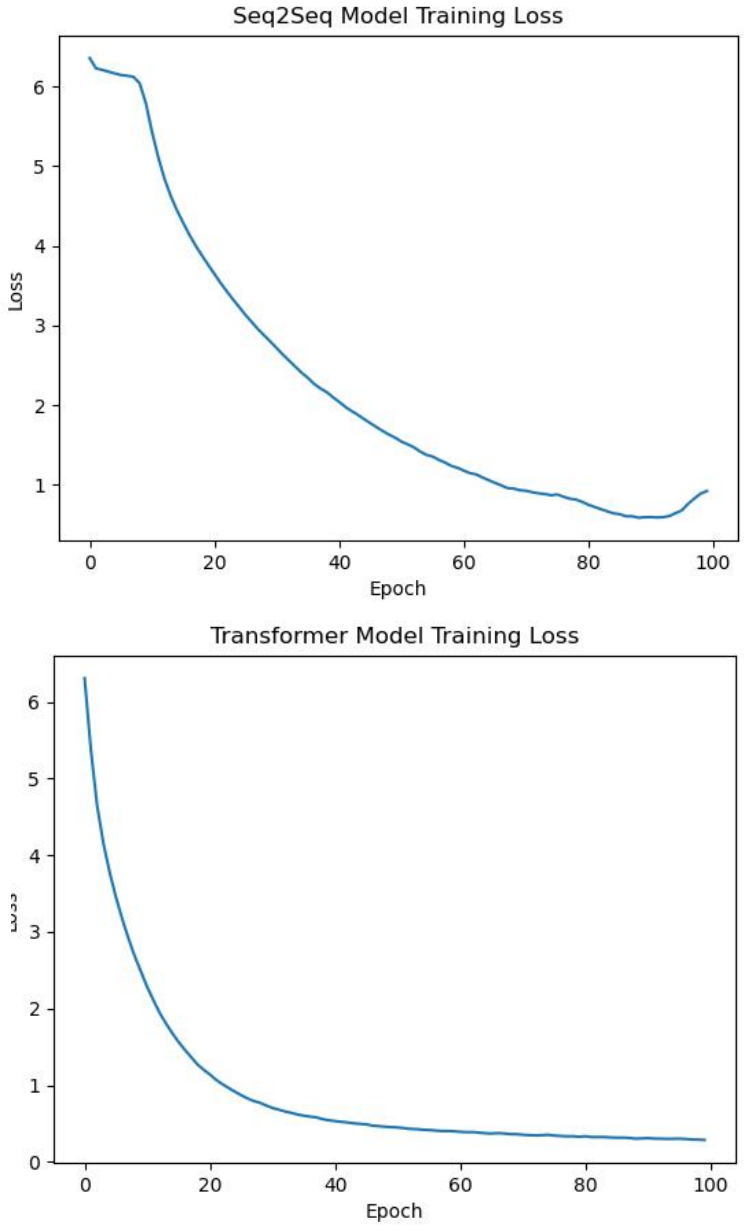

图 3 Seq2Seq 和 Transformer 的训练过程曲线

2. 文本质量

以《雪山飞狐》中的这一句为开头：“田青文接过羽箭，只看了一眼，”，让训练后的模型生成 100 个汉字。

生成结果如表 1 所示。

表 1 模型生成内容

模型	生成内容（含开头）
Seq2Seq	田青文接过羽箭，只看了一眼，叫道：大家随死身。曹云奇话到口边，又咽了过去。曹云奇忍耐不住，急声叫道：误得我在这里干事。你让我给官夫杀了，原来请问师兄，你知你。不住我访什麽？青妹吓著两个僮儿和金面佛的手腕长杰，却没给我师祖爷，果真
Transformer	田青文接过羽箭，只看了一眼，立即抓起一条绳子，捆在苗若兰身上。众人一齐望著刘元鹤身去，只见宝树坐下平主，众人都感愕然跪下。煞时，吮著烈火，他摆弄铁盒，站在阮士中背后，抓住两名侍掌俱放刀刺，身不由主之中，那匕首一个白雪，听得嘿嘿

从表 1 可以看出，两种模型生成的句子都比较符合逻辑。但是 Seq2Seq 所生成的句子之间的关系有点让人摸不到头脑。相比之下，Transformer 生成的内容更为通顺，上下文衔接更自然，验证了 attention 机制的优越性。

Conclusions

本次作业的仿真实验结果表明，两种模型都有给定一段开头生成后续文字内容的能力。相比之下，Seq2Seq 模型由于用了简单的时序 NN，对于长段落上下文的关联掌握的较差。性能上, 其收敛速度较慢, 且文本生成的质量相对一般; Transformer 模型由于运用了 attention 机制，可以无视文本长度、全方位无死角的掌握文本内容相关性，因此性能较好，表现在收敛速度较快，且文本生成质量较高。

综上，Transformer 模型是更好的文本生成模型。

References

[1] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[J]. Advances in neural information processing systems, 2014, 27.

[2] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.