Report of Deep Learning for Natural Language Processing

种雨萌

2775965921@qq.com

Abstract

本篇报告从金庸的小说中选取部分经典作品作为语料库，利用 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），并对比与讨论了这两种方法的优缺点。

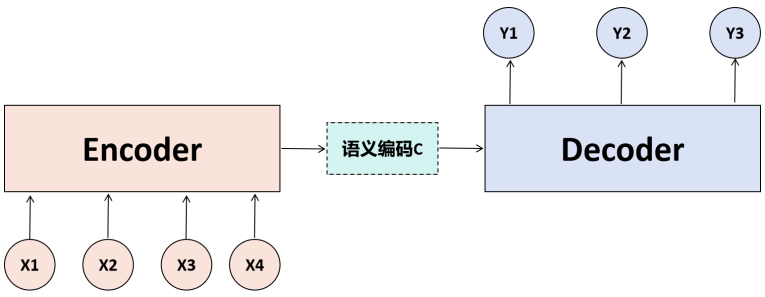
Introduction

Seq2Seq（Sequence to Sequence）和 Transformer 都是用于自然语言处理的重要模型架构，主要用于机器翻译、文本生成等任务。本次实验将利用给定语料库（金庸小说集），通过 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。

Methodology

M1: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，是 RNN 的一个变种，可解决 RNN 要求序列等长的问题。

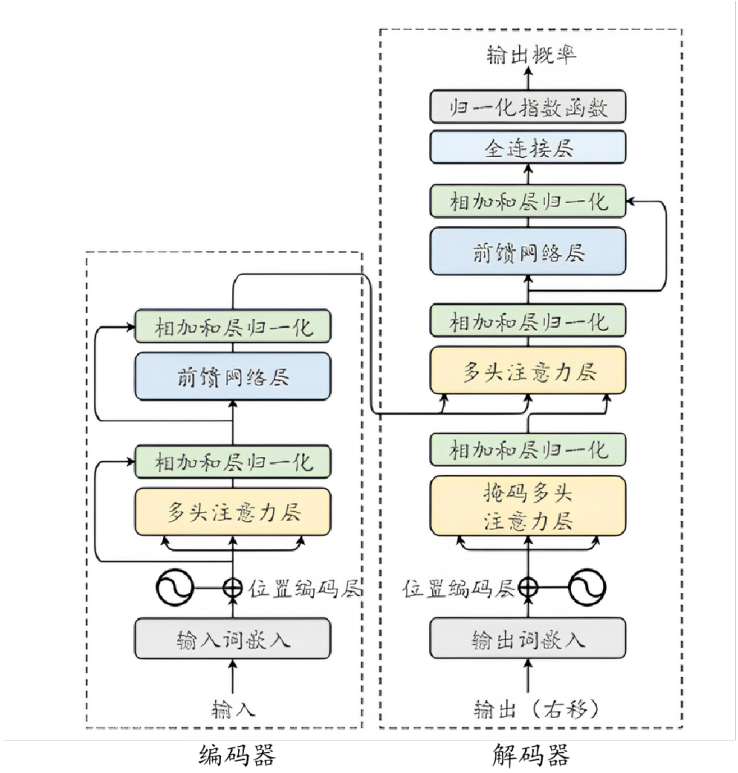


Seq2Seq 最基础的结构如上图所示，在编码过程中，输入序列通过 Encoder，

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。Encoder 和 Decoder 都是 RNN 单元，C 可以看作输入序列内容的一个集合，输入序列所有的语义信息都包含在 C 这个向量里面。

M2: Transformer 模型

Transformer 是一种用于自然语言处理和其他序列到序列任务的深度学习模型架构。Transformer 架构引入了自注意力机制（self-attention mechanism），使其在处理序列数据时表现出色。



如上图所示，Transformer 模型由编码器和解码器模块堆叠在一起，每个编码器层独立处理输入序列，使模型能够学习分层表示并捕获数据中的复杂模式，然后将其输出传递给解码器，后者根据输入生成最终的输出序列。

M3：程序设计

训练文本与测试样本生成：删除文本中所有特殊字符并以句号“。”为界限，将整个文本分割成单独的句子。对于句子的选择，采取以下标准：句子中必须含有字符“她”；句子长度应在 10 到 40 个字符之间；该句的下一句话长度也应在 10 到 40 个字符之间。筛选出 300 句符合上述条件的句子作为训练样本，每个样本的训练目标是其后的一句话。同时额外选取 10 句符合条件且与训练样本不重复的句子作为测试样本。

模型构建：Seq2Seq 模型中的 RNN 均采用 LSTM 模型，编码器和解码器的文字编码嵌入维度均设为 150；编码器和解码器隐藏层维度均设为 100。Transformer 模型中的编码器和解码器的嵌入层维度均设置为 256，隐藏层维度均设置为 512。

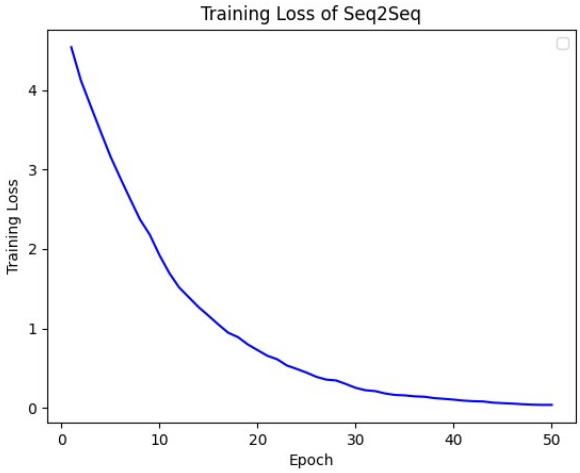
One-hot 字典生成：为处理得到的训练样本和测试样本中的每一个字符进行不重复地编号，从而构建一个 one-hot 编码字典。

批次数据对齐处理：为了在训练过程中统一处理不同长度的输入序列，在每个序列的开始添加一个特殊的开始标识符“<BOS>”，并在序列末尾添加结束标识符“<EOS>”。为了使同一批次中的所有序列长度一致，在每个序列的末尾添加“<PAD>”填充标识符，直到达到该批次中最长序列的长度。

模型训练设置：Seq2Seq 模型和 Transformer 模型均设置迭代训练 50 代，批次大小设置为 2，学习率设置为 0.001。

Experimental Studies

Seq2Seq 模型的训练结果如下图所示。



```
----------------Result 1----------------
Source sentence: 我尽力而为到了这步田地，也已仁至义尽，对得她住
True target sentence: 我立时便走，薛神医能不能救她，只好瞧她的运气了
Generated target sentence: 这一只听得很不耐，一咱们便即回过身来，回复的话了
----------------Result 2----------------
Source sentence: 段誉见她目光中流露恐惧的神气，心想：“王夫人杀人如草芥，确是令人魂飞魄散
True target sentence: ”那少女缓步走到青石凳前，轻轻巧巧的坐了下来，却并不叫段誉也坐
Generated target sentence: 别惊我眼，你得不会去的治
----------------Result 3----------------
Source sentence: ”阿紫无法挖到钟灵的眼珠，便以言语相刺，总是要她大感伤痛，这才快意
True target sentence: 钟灵一听之下，甚是恼怒，但想她这几句话倒也有理，恼怒之情登时变了愁闷
Generated target sentence: 王语嫣道：“你说他是女人？”阿紫道：“当然啦，她身上好香，全是女人的香气
----------------Result 4----------------
Source sentence: 我不出去！我不出去！”她刚才还在大叫“我要出去”，可是一会儿便又大叫“我不出去”
True target sentence: 段誉知她心情激动，一时无可理喻，当下不再说话
Generated target sentence: 当李秋水面前那中年，俯身低声说道：“师叔，师伯有帮手来啦，我背了你逃走
----------------Result 5----------------
Source sentence: 平婆婆滚近木婉清身畔，右手短刀往她小腿上削去
True target sentence: 木婉清飞腿将她踢了个筋斗，就在此时，瑞婆婆的铁拐已点到眉心
Generated target sentence: 到王语嫣肩头是一惊，不会划船她的场

----------------Result 6----------------
Source sentence: 她在对付鸠摩智这贼秃，那是朋友而非敌人
True target sentence: ”便道：“老夫人尽可放心，在下既到尊府，一切但凭老夫人吩咐便是
Generated target sentence: 王语嫣道：“是啊，你好写了，我这才死得瞑目
----------------Result 7----------------
Source sentence: ”朱丹臣道：“快去找那婆子，别让她走了
True target sentence: ”木婉清奔向厨房，巴朱二人追出木屋
Generated target sentence: ”四条汉子应声跃起，分从两侧包抄了上来
----------------Result 8----------------
Source sentence: ”她拍掌两下，小茗了过来
True target sentence: 王夫人道：“你传下话去，有谁和那姓段的花匠多说一句话，两人一齐都割了舌头
Generated target sentence: 当然老妈妈煎一碗姜汤给你喝
----------------Result 9----------------
Source sentence: 只听段正淳道：“那么咱们去问你师妹，她一定知道誉儿关在什么地方
True target sentence: ”刀白凤怒道：“不许你去见甘宝宝
Generated target sentence: 不料我听缘根说，你好不如此，就怕你觉得不好
----------------Result 10----------------
Source sentence: 这女娃儿似乎是个丫鬟之类，她突然抬头，我一个闪避不及，跟她打了个照面
True target sentence: 在下深恐泄露了机密，纵上前去，施展擒拿法，便想将她抓住
Generated target sentence: ”一红色道：“你……你何必吓我？”阿朱道：“我不是吓你
```

Transformer 模型的训练结果如下图所示。

```
----------Result 1----------
Source sentence: 阮星竹一见，脱口叫道：“阿紫！”她忘了自己改穿男装，这一声叫，是本来的女子声音
True target sentence: 右首马上乘客身穿百结锦袍，脸上神色木然，俨如僵尸
Generated target sentence:
----------Result 2----------
Source sentence: ”段正淳道：“她独自常常使这掌法？”木婉清点头道：“是
True target sentence: 师父每次练了这套掌法，便要发脾气骂我
Generated target sentence: 多不之笑那，个也一，轻，
----------Result 3----------
Source sentence: 我抱你去还给你主人，她一定喜欢得不得了
True target sentence: ”学着钟灵吹口哨的声音，嘘溜溜的吹了几下
Generated target sentence: 得不
----------Result 4----------
Source sentence: ”她一直犹豫难决，刚才一场变故却帮她下了决心
True target sentence: 阿朱喜道：“姑娘肯去援手，当真再好也没有了
Generated target sentence: ，“道包
----------Result 5----------
Source sentence: 段正淳抢上去将她搂住
True target sentence: 甘宝宝身子一颤，晕了过去
Generated target sentence: ，段人的那大，”中，了老，的理，到

----------Result 6----------
Source sentence: ”她举手便即杀人，自也不怕什么死人，好奇心起，快步走过去察看
True target sentence: 见这青袍人是个老者，长须垂胸，面目漆黑，一双眼睁大大的，望着江心，一霎也不霎
Generated target sentence: 的，，然当段
----------Result 7----------
Source sentence: ”童姥道：“她虽知道我进了皇宫，却不知我躲在何处
True target sentence: 皇宫中房舍千百，她一间间的搜去，十天半月，也未必能搜得到这儿
Generated target sentence: 见上
----------Result 8----------
Source sentence: 段誉仍在催问阿紫，她明日和王语嫣约定在何处相见
True target sentence: 阿紫见他如此情急，心下盘算如何戏弄他一番，说不定还可捡些便宜，当下只是顺口敷衍
Generated target sentence: 的我带李这，上，可“”，，，，，了那
----------Result 9----------
Source sentence: 她母亲已给人点了穴道，却动弹不得
True target sentence: 过不多久，段正淳手下有五六个人到来
Generated target sentence:
----------Result 10----------
Source sentence: 她不置可否，慢慢低下头来，眼睛中流露出异样的光彩
True target sentence: 邓百川和公冶乾对望了一下，觉得欺骗了这个天真烂漫的姑娘，心中颇感内咎
Generated target sentence: 只，的，”不
```

首先对比两种模型下损失值的变化趋势。可以看到 Seq2Seq 模型下的损失值快速下降，说明 Seq2Seq 模型迅速收敛；而 Transformer 模型下的损失值始终维持在较高的区域内，说明 Transformer 模型很难收敛。

然后对比两种模型下文本生成的效果。可以看到 Seq2Seq 模型下生成的文本语意较为完整，虽然与原句不符，但与上一句还存在一定逻辑关系；而 Transformer 模型下生成的文本基本都是一两个字的短词，几乎不成整句，毫无语意可言，更不用提与上一句是否存在逻辑关系。通过以上分析，可以总结得到 Seq2Seq 模型在处理较短序列或小规模数据集的能力可能强于 Transformer 模型。

最后分析两种模型的优缺点。Seq2Seq 模型简单、有效，适合处理较短序列或小规模数据集，但在处理长序列和并行计算方面存在局限。Transformer 模型具有强大的表现力和并行计算能力，适合处理长序列和大规模数据，但需要更多的计算资源和数据量，模型复杂度较高。

Conclusion

本篇报告从金庸的小说中选取部分经典作品作为语料库，利用 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），并对比与讨论了这两种方法的优缺点。