NLP 作业

——使用 Seq2Seq 与 Transformer 模型进行文本生成的比较研究

院（系）名称	自动化科学与电气工程学院
学生姓名	胡正皓
学生学号	ZY2303205

2024 年 6 月 16 日

一、引言

在自然语言处理（NLP）领域，文本生成是一项重要的任务。通过给定文本的开头，生成连贯且风格一致的文本片段或章节，可以应用于多种场景，如自动写作、对话系统等。本研究旨在利用金庸小说的语料库，分别采用 Seq2Seq 和 Transformer 模型来实现文本生成，并对比两种方法的优缺点。

二、研究方法

语料库的选择：本研究采用的中文语料库由金庸撰写的武侠小说越女剑。如果加载全部的小说，训练时间过长，因此选了一本最短的小说，作为语料库。

数据处理方法：鉴于原始文本中含有大量的乱码、无效内容以及重复的中英文符号，对数据集进行彻底的预处理变得尤为重要。预处理步骤包括：首先，去除所有隐藏的字符以清理文本；其次，删除所有非中文字符，确保分析的纯净度；最后，在不考虑上下文的情况下，移除所有标点符号，避免对分词结果产生干扰。

本研究采用了 jieba 分词库进行文本处理，jieba 是 Python 中广泛使用的一个中文分词工具。在此实验中，我们采用了 jieba 的精确模式进行分词，旨在最大程度上保证文本分词的准确性和效率。这一系列预处理措施为后续的数据分析提供了干净、可靠的文本基础。

模型构建与训练：

Seq2Seq 模型：

Seq2Seq（Sequence to Sequence）模型是一种常用于机器翻译、文本生成等任务的神经网络架构。Seq2Seq 模型的核心思想是使用一个编码器（Encoder）将输入序列编码为一个固定长度的上下文向量，然后使用一个解码器（Decoder）将这个上下文向量解码为输出序列。

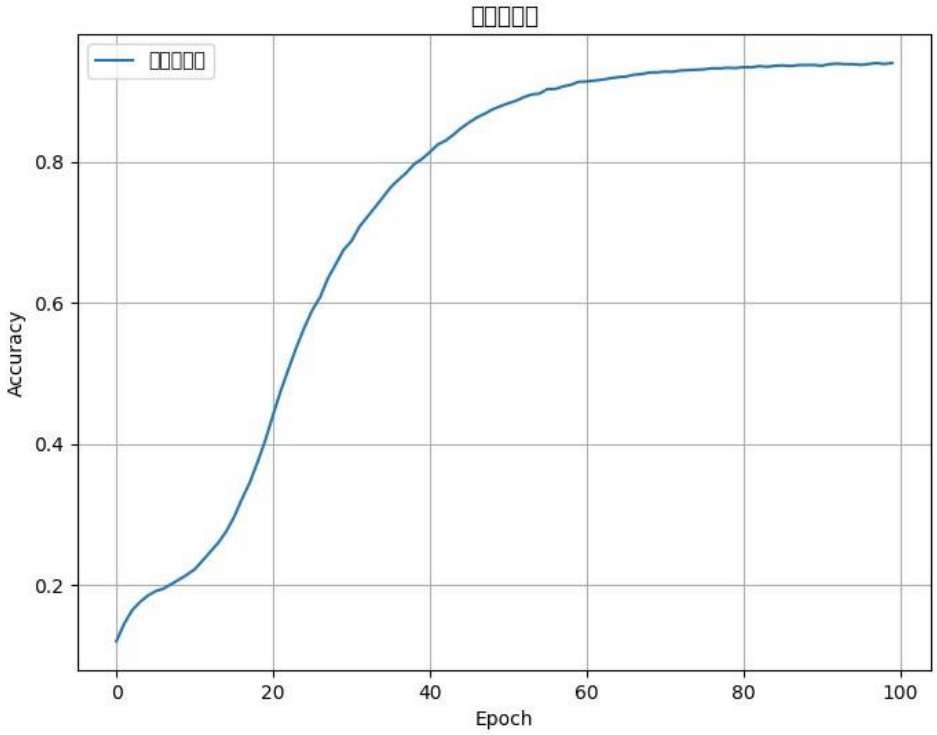
编码器（Encoder）：

编码器通常由一个 RNN（如 LSTM 或 GRU）构成，它将输入序列逐步处理，并将每个时间步的隐藏状态传递给下一个时间步。最后一个时间步的隐藏状态被视为整个输入序列的表示。

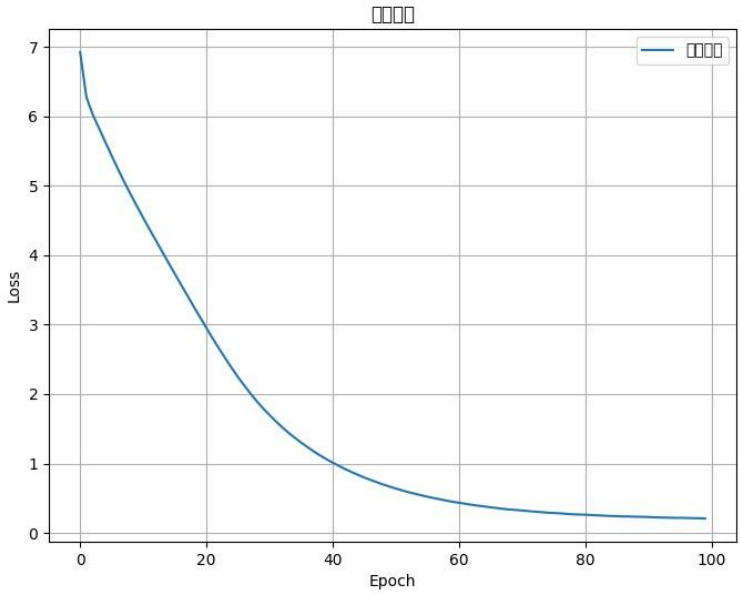
解码器（Decoder）：

解码器也是一个 RNN，接收编码器的输出（即上下文向量）作为其初始状态。解码器在每个时间步生成一个输出，并将其作为下一个时间步的输入，直到生成结束符。

下面是模型的训练曲线：



正确率曲线

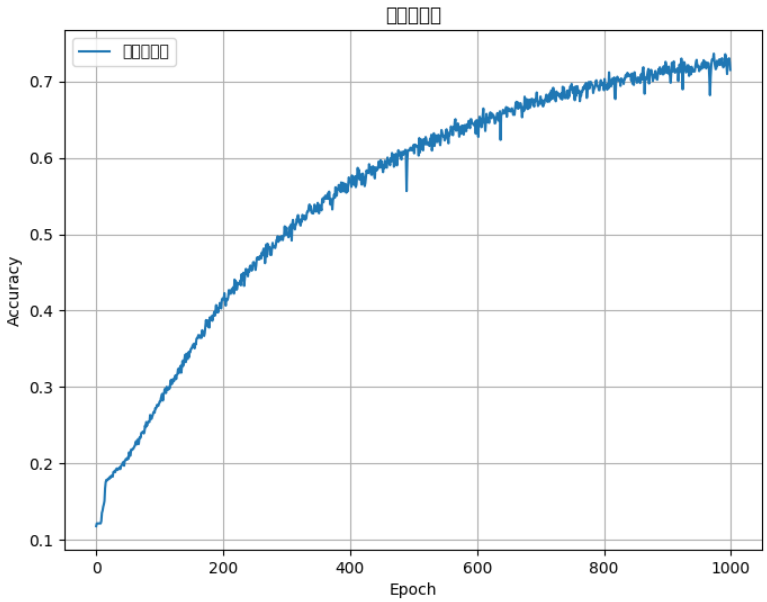


损失函数曲线

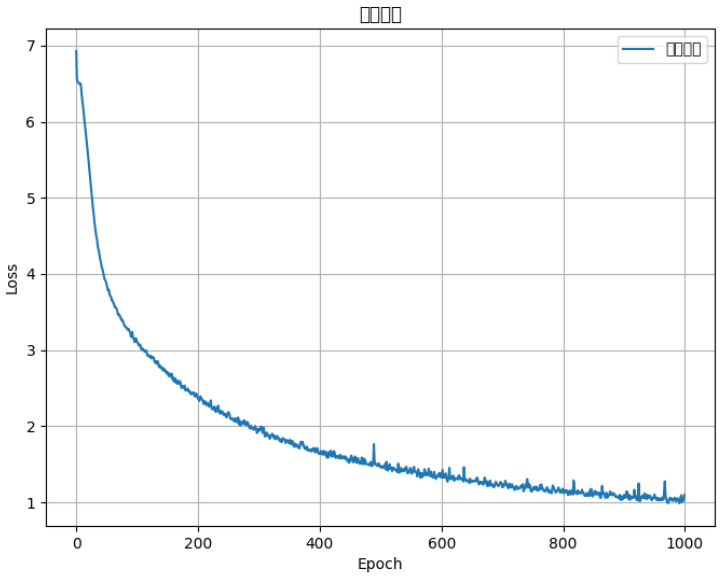
Transformer 模型：

Transformer 模型是一种基于自注意力机制的神经网络架构，广泛应用于 NLP 任务。与传统的 Seq2Seq 模型不同，Transformer 模型不使用循环结构，而是通过多头自注意力机制并行处理整个序列，

从而提高训练效率和效果。Transformer 模型的核心组件包括编码器和解码器，每个组件由多个相同的层堆叠而成，每层包含一个多头自注意力机制和一个前馈神经网络。



正确率曲线



损失函数曲线

结果分析

原文	Seq2seq	原文	Transformer
这时场中两名青衣剑士仍以守势缠住了一名锦衫剑士，另外	这时场中两名青衣剑士仍以守势缠住了一名锦衫剑士，另外两名青衣剑士快剑攻击，杀死第三名锦衫剑士后，转而向第四名敌手相攻至小腹，划了一道两尺来长的口子，心中便已能分别小人二，震骇之下，心中	一名吴士兴犹未尽，长剑一挥，将一头山羊从头至臀	一名吴士兴犹未尽，长剑一挥，将一头山羊从头至臀，一道一道，颤动下难破越越越越身子贯入贯入大笑，，大笑贯入替越天真，，替越一听替越替越，天真吴兵天真，几个便装挡格几个几个便装便装纵跃

三、结论

通过对比 Seq2Seq 和 Transformer 模型在金庸小说语料库上的文本生成表现，我们可以看到两种模型各有优缺点。Seq2Seq 模型简单易懂，适合较短文本生成，而 Transformer 模型在处理长距离依赖和生成连贯文本方面表现更优。未来的研究可以进一步优化模型和数据预处理方法，以提高文本生成质量。