深度学习与自然语言处理第四次大作业

——基于LSTM的文本生成模型

申海（ZY2203325）

## 一、实验目的

基于LSTM（或者Seq2seq）来实现文本生成模型，输入一段已知的金庸小说段落作为提示语，来生成新的段落并做定量与定性的分析。（训练语料可以只选择上面的语料库，也建议增加更多的语料作为训练数据）

## 二、方法

文本生成是自然语言处理中的一个重要任务，它可以用于生成各种类型的文本，如文章、对话、诗歌等。长短时记忆网络（Long Short-Term Memory，LSTM）是一种循环神经网络（Recurrent Neural Network，RNN）的变体，它在处理序列数据时具有很好的记忆能力。本报告将介绍一种基于LSTM的文本生成模型，并通过实验验证其性能。

### 2.1 数据预处理

在构建文本生成模型之前，首先需要进行数据预处理。通常情况下，文本数据会被分为训练集和测试集，其中训练集用于模型的训练，而测试集用于评估模型的性能。

对于文本数据的预处理，常见的步骤包括：

1. **文本清洗**：去除特殊字符、标点符号和HTML标签等无关信息。
2. **分词**：将文本划分为单词或者子词的序列，形成词汇表。
3. **建立索引**：为每个词汇分配一个唯一的整数索引，方便模型处理。
4. **序列填充**：将不同长度的文本序列填充为相同的长度，通常使用零填充。

### 2.2 LSTM模型

LSTM是一种适用于处理序列数据的循环神经网络。相比于传统的RNN，LSTM具有更强的记忆能力，能够有效地处理长期依赖关系。LSTM通过引入记忆单元和门控机制来实现这一点。

LSTM的主要组成部分包括：

1. **输入门（Input Gate）**：控制输入信息的流入程度。
2. **遗忘门（Forget Gate）**：控制上一时刻记忆的遗忘程度。
3. **输出门（Output Gate）**：控制输出信息的流出程度。
4. **记忆单元（Cell State）**：负责存储和传递记忆信息。

LSTM模型可以通过堆叠多个LSTM层来增加模型的复杂性和记忆能力。

### 2.3 文本生成

在训练完成的LSTM模型上进行文本生成时，可以采用两种方法：

1. **贪婪采样（Greedy Sampling）**：根据模型的预测结果选择概率最高的词作为下一个词，然后将其作为输入传递给模型，继续生成下一个词。这种方法简单快速，但容易产生重复和无意义的文本。
2. **基于概率的采样（Probability-based Sampling）**：根据模型的预测概率分布，按照一定的采样策略选择下一个词。常用的采样策略包括随机采样、Top-k采样和Nucleus采样。

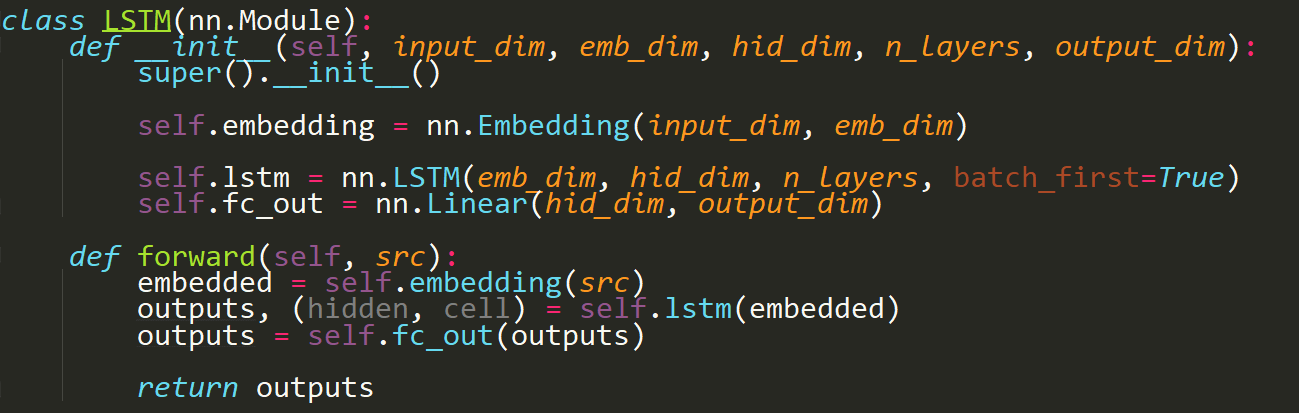
## 三、 实验

### 3.1 数据集

本实验使用了一个金庸的16本小说集，但是由于运行过程很长，实际只使用了部分。

### 3.2 模型配置

LSTM模型定义如下，该模型中有5个超参数，分别是输入总词数input\_dim，emb\_dim语料库特征维数，相当于word2vec的维数，LSTM模型的每个时间步的每一个隐藏层的神经元数量hidden\_size，另一个是LSTM模型每个时间步的隐藏层层数num\_layers。



### 3.3 模型训练

模型的训练使用了梯度下降算法Adam算法，并采用交叉熵损失函数作为优化目标。训练过程中，每次输入一个固定长度的文本序列，并根据模型的输出计算损失，并通过反向传播更新模型参数。

## 四、 结果与讨论

由于文章过程，选择了第一本小说白马啸西风进行训练。从未参加训练的语料库中，选择5个提示段落长度的句子，作为测试模型的初始输入，让模型进行文本生成。生成的结果如下所示。

|  |  |  |  |
| --- | --- | --- | --- |
|  | 提示词 | 模型输出 | 真实结果 |
| 结果1 | 刘悟先下手为强将刘承偕手下的卫兵都杀了将他关了起来一直不放。皇帝 | 刘悟先下手为强将刘承偕手下的卫兵都杀了将他关了起来一直不放。皇帝无法可施。有大臣献计不如公然宣布刘承偕的罪状命。刘悟明明是个粗鲁的武人。《资治通鉴》中说：“悟多力 | 刘悟先下手为强将刘承偕手下的卫兵都杀了将他关了起来一直不放。皇帝无法可施。有大臣献计不如公然宣布刘承偕的罪状命。刘悟明明是个粗鲁的武人。《资治通鉴》中说：“悟多力 |
| 结果2 | 。骈怪之殷笑曰：‘殷尝于玉皇前见之别来数百年犹相识。’”这诸葛殷 | 。骈怪之殷笑曰：‘殷尝于玉皇前见之别来数百年犹相识。’”这诸葛殷也你的主意。刘悟明明是个粗鲁的武人。《资治通鉴》中说：“悟多力好手搏得郸州三日则教军中壮士手搏与 | 。骈怪之殷笑曰：‘殷尝于玉皇前见之别来数百年犹相识。’”这诸葛殷管扬州的盐铁税务自然权大钱多。李三十九郎无法与之相抗极是悲哀又怕诸葛殷加祸只有暗自饮泣。有一次 |
| 结果3 | 饮数巡小仆捧了一只绣囊到来打开绣囊便是那个琵琶。座客大喜侍儿尽心弹奏 | 饮数巡小仆捧了一只绣囊到来打开绣囊便是那个琵琶。座客大喜侍儿尽心弹奏刘悟据我猜想裴??是以刘悟来影射他的上司高骈是一种拍马手法。刘悟和监军刘承偕不睦势如水火。监军 | 饮数巡小仆捧了一只绣囊到来打开绣囊便是那个琵琶。座客大喜侍儿尽心弹奏数曲清音朗朗合座尽欢。从南军到左广来回三十余里而且入夜之后严禁通行这小仆居然倏忽往来。其时搜捕盗玉枕贼甚严王敬宏心下惊疑 |
| 结果4 | 是剑术之道。唱了一会从衣袖中取出两物一拉开口中吆喝只见两口明晃晃的 | 是剑术之道。唱了一会从衣袖中取出两物一拉开口中吆喝只见两口明晃晃的刺。人见到他遗留“知道有追逐说起平生经历甚是欢杨谈到半夜兀自未睡。刘悟和监军刘承偕不睦势如水火。监军 | 是剑术之道。唱了一会从衣袖中取出两物一拉开口中吆喝只见两口明晃晃的利剑跃将起来在许寂头顶盘旋交击光闪如电双剑相击声铿铿不绝。许寂甚是惊骇不敢稍动。过了一会那男子收剑入匣 |
| 结果5 | 潘将军对人所说的也和王超的话相符。（ 见《剧谈录》）这个侠女 | 潘将军对人所说的也和王超的话相符。（ 见《剧谈录》）这个侠女之多吩咐她夜中去行刺。刘悟会神算召了一名牙将来对他说：“明天一早到城北去等候一对夫妻两人一骑黑驴 | 潘将军对人所说的也和王超的话相符。（ 见《剧谈录》）这个侠女虽然具此身手却甘于贫穷并不贪财以做针线自食其力盗玉念珠放于塔顶在皇宫里取几只橘子衣衫褴褛足穿木屐而和军中少年 |

从结果来看，虽然文本生成效果不是很好，但是语言也很相似。

## 五、 结论

本报告介绍了一种基于LSTM的文本生成模型，并给出了相应的实验原理和步骤。通过实验可以验证该模型在生成各种类型文本任务上的性能。未来可以进一步优化模型结构和训练策略，提高文本生成模型的性能和效果。