单位代码_____
学　　号 ZY2203811
分 类 号_____

北京航空航天大學
BEIHANG UNIVERSITY

基于 Seq2Seq 模型的金庸小说数据集的文本生成

院（系）名称 自动化科学与电气工程学院

专业名称 自动化

学生姓名 许铁

2023 年 05 月

1 内容介绍

自然语言处理的核心便是为语言建立合理的数学模型，既而探究文本的结构，相当于将抽象的语言映射到了一个清晰的数学系统，那么，应用这个数学系统我们便可以进行文本分类、文本生成等工作。在文本分类领域 Seq2Seq 模型有着举足轻重的地位。解决很多模型受限制的输出只是一个参量的问题，完成输入和输出均为序列的问题。该模型简单有效，对任务具有良好的泛化性本篇报告以金庸小说为文本语料，实现基于 Seq2Seq 模型的文本生成。

1.1 实验要求

基于 Seq2Seq 模型来实现文本生成的模型，输入可以为一段已知的金庸小说 段落，来生成新的段落并做分析。

1.2 数据集介绍

本次数据集为 16 本金庸武侠小说，武侠小说辞藻华艳而又通俗易懂，贴近生活而又包罗万象， 文白夹杂而又雅俗共赏， 非常适合作为本次实验的数据集。在所有的武侠小说作家中，金庸的文笔首屈一指，语言流畅、凝练、准确、画面感强。本次数据包含 “飞雪连天射白鹿，笑书神侠倚碧鸳”十四篇长篇小说以及《越女剑》和《三十三剑客图》，基本上涵盖了金庸的所有武侠作品。

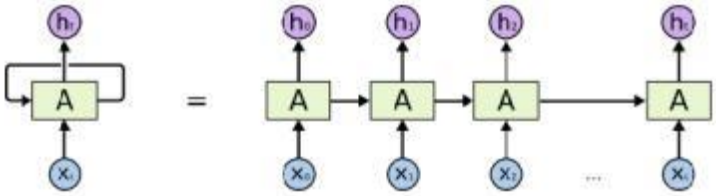
数据库地址： https://share.weiyun.com/5zGPyJX

2 实验原理

2.1 RNN 和 LSTM

2.1.1 基本概念

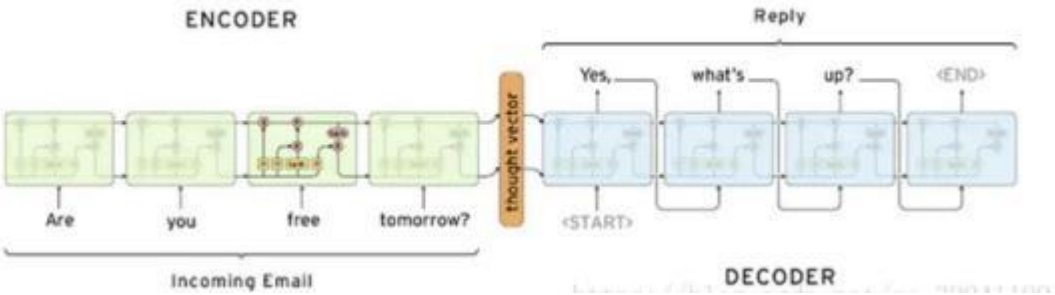
传统的 RNN 模型是一种节点定向连接成环的人工神经网络，是一种反馈神经网络，RNN 利用内部的记忆来处理任意时序的输入序列，并且在其处理单元之间既有内部的反馈连接又有前馈连接，这使得 RNN 可以更加容易处理 不分段的文本等。但是由于 RNN 存在梯度消失问题，无法“记忆”长时间序列 上的信息，只能对部分序列进行记忆，所以在长序列上表现远不如短序列，造 成了一旦序列过长便使得准确率下降的结果。而 LSTM 的提出一定程度上解 决了这一问题。在标准的 RNN 模型中，其链式形式的结构模块中只有一个简 单的结构。其信息传递流程如图所示。



LSTM(long short term memory，长短期记忆)模型，是一种特殊的 RNN 模型。相较于 RNN 模型，LSTM 模型增加了输入门、输出门、忘记门三个控制单元，随着信息的进入该模型，LSTM 会对信息进行判断，符合规则的信息会被留下，不符合的信息会被遗忘，以此原理，可以解决神经网络中长序列依赖问题。LSTM 模型中 t 时刻下的第一步是决定丢弃 ht-1 与 xt 中的部分信息，通过忘记门来完成。下一步是保存部分信息，将新的信息选择性的记录到状态中，通过输入门来完成。最后一步是确定输出值，通过输出门确定。

2.2 Seq2Seq 模型

目前 Seq2Seq 模型在机器翻译，语音识别，文本摘要，问答系统等领域取得 了巨大的成功。Seq2Seq 属于 encoder-decoder 结构的一种，这里看看常见的 encoder-decoder 结构，基本思想就是利用两个 RNN，一个 RNN 作为 encoder， 另一个 RNN 作为 decoder。encoder 负责将输入序列压缩成指定长度的向量，这 个向量就可以看成是这个序列的语义，这个过程称为编码，如下图，获取语义向 量最简单的方式就是直接将最后一个输入的隐状态作为语义向量 C。也可以对最 后一个隐含状态做一个变换得到语义向量，还可以将输入序列的所有隐含状态做 一个变换得到语义变量。而 decoder 则负责根据语义向量生成指定的序列，这个 过程也称为解码，如下图，最简单的方式是将 encoder 得到的语义变量作为初始 状态输入到 decoder 的 RNN 中，得到输出序列。



Attention 机制:由于基础 Seq2Seq 模型的上述缺陷，随后引入了 Attention 的概念，Attention 在 decoder 过程中的每一步，都会给出每个 encoder 输出的 特定权重，然后根据得到权重加权求和，从得到一个上下文向量，这个上下文向 量参与到 decoder 的输出中，这样大大减少了上文信息的损失，能够取得更好的 表现。

3 实验过程

本次实验利用 Seq2Seq 模型进得到文本生成模型，在 Seq2Seq 模型中的 Encoder 和 Decoder 模块都利用 LSTM 模型进行训练。训练时利用金庸先生的 16 本小说作为实验数据集，进行文本生成模型的训练和测试实验。

3.1 数据预处理

由于数据库里存在各种标点符号以及网页信息，所以首先需要对数据进行预 处理操作。

删除 txt 文件中关于网址描述的与金庸武侠小说内容无关的字符"本书来自 www.cr173.com 免费 txt 小说下载站\n 更多更新免费电子书请关注 www.cr173.com","本书来自 www.cr173.com 免费 txt 小说下载站

删除非中文字符，根据中文字符的 utf-8 编码的字节长度为 3 来判断；

删除标点符号，并且根据带有分割意义的标点符号['\n',' 。',' ? ',' ! ',' ， ',';',' ：',' . ']对文本进行按句换行分割。

3.2 分词

本文选择"结巴（jieba）"中文分词模块，该模块可以支持三种分词模式：精确模式，试图将句子最精确地切开，适合文本分析；全模式，把句子中所有的可以成词的词语都扫描出来,速度非常快，但是不能解决歧义；搜索引擎模式，在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。同时，由于金庸小说中包括部分繁体字，该模块可以支持繁体分词、支持自定义词典。

本文使用jieba.cut()进行分词， 例如对以下一句话：

"武林至尊宝刀屠龙号令天下莫敢不从倚天不出谁与争锋"

进行断句后得到：

['武林','至尊','宝刀','屠龙','号令','天下','莫敢','不','从','倚天','不出','谁','与','争锋']

可以看出jieba 可以对中文句子很好地进行分词操作。之后进行数据集制作。

3.3 训练模型

（1）字典生成

将文本语料库 corpus_chars 的字符不重复统计，可以得到一个字典， 并且给字典的每个字符对应一个索引，本来语料库是由中文字符组成的，可以通过字典来将字符转换成索引，得到索引 组成的语料库

（2）Word Embedding

建立字典可以将字符变成索引，还需将索引变成词向量，这一步叫做词嵌入， 即 Word Embedding，词向量可以是不用训练的，比如 one-hot，也可以是需要训 练的，比如使用 torch.nn.Embedding()。本次实验使用 one-hot 向量。

（3）数据集生成

num_steps, batch_size 两个参数分别代表训练集的文本序列长度和批样本数量。输入进网络的文本可以表示成[batch_size,num_steps]的一个索引 tensor。这一步通过对 corpus_indices 切片分块来实现，前 num_steps 个 token 作为输入，后 num_steps 个 token 作为输出。

（4）seq2seq 模型

在本次的 seq2seq 模型中，编码器和解码器都是采用 LSTM 网络，直接使用 pytorch 的 torch. nn.LSTM(input_size, hidden_size,num_layers)模块。

input_size 代表输入 sequence 的特征维度；

hidden_size 代表 hidden state 的特征维度；

num_layers 代表 LSTM 网络层数。

由于输入的是 one-hot 向量， 维度为字典长度 len(char_to_idx)=1186，hidden_size 可以设置为 128，256，512，1024。num_layers 可以设置为 1，2 等。

loss = nn.CrossEntropyLoss()

optimizer = torch.optim.Adam(model.parameters(), lr=lr)
反向传播过程中使用了梯度裁剪 grad_clipping()

4 实验结果与分析

4.1 实验结果

样例	输入	输出
1	看敌人时，当先一人身形瘦削，脸色漆黑，认得是北京平通镖局总镖头熊元献，此人精熟地堂刀功夫。饮马川山寨曾劫过他镖局的一支大镖，熊元献使尽心机，始终没能要回，双方结下甚深梁子。另一个女子三十二三岁年纪，马寨主识得她是双刀郑三娘。	这位熊元献道： 咱们重新胆雅，小小说吴叁桂演义和铁冠图叙述李自成故事， 只怕你一直孤苦伶仃，还 是这雪峰。宝树道： 这位大英雄的朋 友， 敝上却有人邀。宝树道: 这位大 英雄的朋友， 敝上却有人邀。殷吉道：咱们快去寻宝， 这一个不敢再说。
2	范蠡摆了摆手，众卫士退了下 去。范蠡握着西施的手， 道： “咱们换 上庶民的衣衫，我和你到太湖划船 去， 再也不回来了。”西施眼中闪出无 比快乐的光芒， 忽然之间，微微蹙起 了眉头， 伸手捧着心口。阿青这一棒 虽然没戳中她， 但棒端发出的劲气已 刺伤了她心口。	阿青道： “范蠡， 你要杀你的西施， 她 逃不了故乡的风光，也不知难而退。” 阿青道：“她家里有大草来吗？那好 前，我要说你是西施。”范蠡又惊又喜 道： “你的嘴唇比这朵小红白，一个，不许卫士们回甚么。
3	石破天心想：石庄主夫妇胯下坐骑 奔行甚快， 我还是尽速赶上前去的为 是。”看明了石清夫妇的去路，跃下树来，从山坡旁追将上去。还没奔过 上清观的观门，只听得有人喝道： “是 谁？站住了！”	石清夫妇心想： “这些泥人儿都是哑 乖，不是我的孙女婿，你们要杀他，我便不是我的天哥，我是狗杂种，怎 么忽然抽筋，不是我的天哥， 我是不 是？”石破天道： “是啊，我不是我的 孙女婿， 你也不是我的。

4.2 定量定性分析

1. num_layers 网络层数

SET：

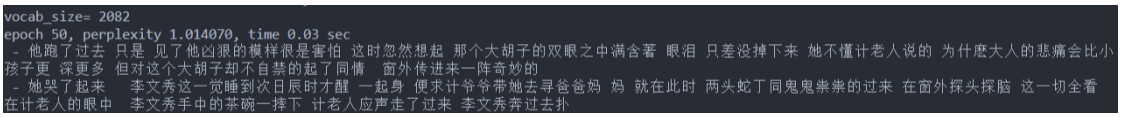
corpus_chars = corpus_chars[0:10000] ，hidden_size=256

num_epochs, num_steps, batch_size, lr, clipping_theta = 50, 35, 32, 1e2, 1e-2

prefixes = ['他跑了过去','她哭了起来'] （输入文本），

pred_len = 70（生成文本长度）

num_layers = 1:



num_layers=2:

分析：

可以看到层数增加，perplexity 在增高，训练时间变大，文本生成结果变差。由于我们的数据集很小因此，一层网络就足够。如果把语料库 10000 字符扩大，可以用更深层的网络来训练。

2. hidden_size 隐藏层的特征维度

SET:

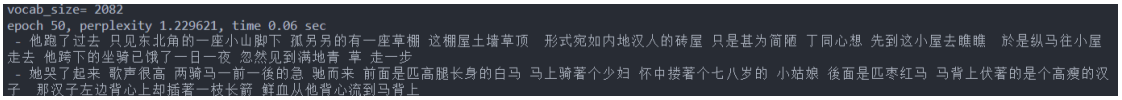
corpus_chars = corpus_chars[0:10000] ，num_layers=1

num_epochs, num_steps, batch_size, lr, clipping_theta= 50, 35, 32, 1e2, 1e-2, 1

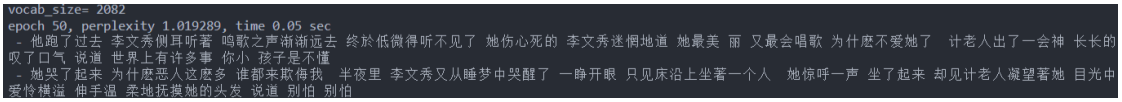
prefixes = ['他跑了过去','她哭了起来'] （输入文本），

pred_len = 70（生成文本长度）

hidden_size=128:



hidden_size = 512：



hidden_size=1024:

分析：

可以看到当 hidden_size=256 和 512 时 perplexity 都较低，文本生成结果都很好。因此隐藏层维度也不是越高越好，都是需要调参的。

3. 是否使用 jieba 分词

SET:

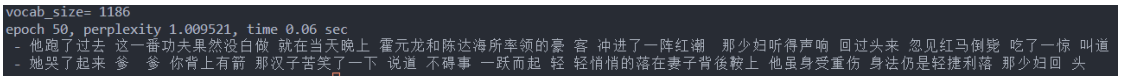
corpus_chars = corpus_chars[0:10000] ，num_layers=1,hidden_size=256

num_epochs, num_steps, batch_size, lr, clipping_theta= 50, 35, 32, 1e2, 1e-2, 1

prefixes = ['他跑了过去','她哭了起来'] （输入文本），

pred_len = 70（生成文本长度）

不适用jieba 分词：



分析：

可以看到不使用jieba 分词，字典的字符数量减小了，但是 perplexity 也很小，文本生成结果也比较好。

5 总结

本文基于 Seq2Seq 模型来实现文本生成的模型，输入可以为一段已知的金庸小说段落，来生成新的段落并做分析。自然语言是一种上下文相关的信息表达和传递的方式，让计算机处理自然语言，一个基本的问题就是为自然语言这种上下文相关的特性建立数学模型，即统计语言模型。自然语言处理的核心便是为语言建立合理的数学模型，既而探究文本的结构，相当于将抽象的语言映射到了一个清晰的数学系统，那么，应用这个数学系统我们便可以进行文本分类、文本生成等工作。在文本分类领域 Seq2Seq 模型有着举足轻重的地位。经过本次作业与课堂学习，我对 Seq2Seq 模型有了全局的认识，系统地学习了其原理，通过学习，对 Seq2Seq 模型生成文档的原理有了深刻的理解。此次作业应用词向量进行文本生成，加深了我们对 Seq2Seq 模型如何进行训练。

6 参考文献

https://blog.csdn.net/weixin_50891266/article/details/116750204

https://github.com/Outlande/Coursework_nlp/