作业4 基于LSTM的文本生成模型

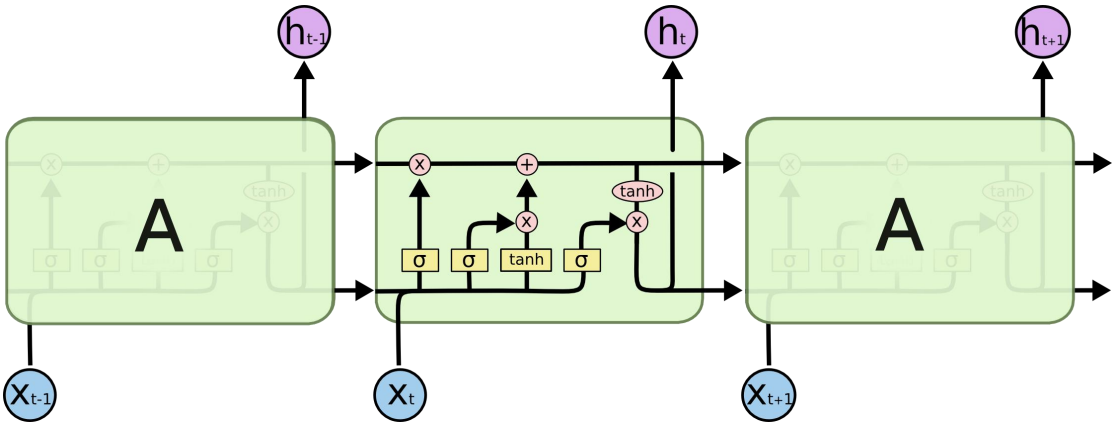
Abstract

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）

Introduction

长短期记忆网络（LSTM，Long Short-Term Memory）是一种时间循环神经网络，是为了解决一般的RNN（循环神经网络）存在的长期依赖问题而专门设计出来的，适合于处理和预测时间序列中间隔和延迟非常长的重要事件。

所有RNN都具有一种重复神经网络模块的链式的形式。在标准的RNN中，这个重复的模块只有一个非常简单的结构，例如一个tanh层。激活函数 Tanh 作用在于帮助调节流经网络的值，使得数值始终限制在 -1 和 1 之间。LSTM同样是这样的结构，但是重复的模块拥有一个不同的结构。具体来说，RNN是重复单一的神经网络层，LSTM中的重复模块则包含四个交互的层，三个Sigmoid 和一个tanh层，并以一种非常特殊的方式进行交互。



上图中， σ 表示的Sigmoid 激活函数与 tanh 函数类似，不同之处在于 sigmoid 是把值压缩到0~1 之间而不是 -1~1 之间。这样的设置有助于更新或忘记信息：

因为任何数乘以 0 都得 0，这部分信息就会剔除掉；

同样的，任何数乘以 1 都得到它本身，这部分信息就会完美地保存下来

相当于要么是1则记住，要么是0则忘掉，所以还是这个原则：因记忆能力有限，记住重要的，忘记无关紧要的。

LSTM的主要组成部分有忘记门、输入门、细胞状态、输出门。忘记门读取上一个输出h_{t-1}和当前输入x_t，经过sigmoid激活后输出向量f_t，该向量的所有值都在0和1之间，1表示完全保留，0表示完全丢弃，之后与上一个细胞状态C_{t-1}相乘。

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

输入门确定什么样的新信息会被存放在细胞状态中，首先sigmoid决定哪些值需要更新，之后tanh创建一个候选值向量C_t加入到细胞状态中。

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

细胞状态更新时，首先旧状态C_{t-1}与f_t相乘，确定需要舍弃的信息，再加上i_t*C_t得到新的候选值C_t。

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

输出门决定要输出的值，输出基于细胞状态，首先通过sigmoid决定将细胞状态的哪部分输出，再将细胞状态通过tanh处理，之后与sigmoid后的结果相乘得到输出部分。

$$o_t = \sigma\left(W_o \; [h_{t-1}, x_t] \; + \; b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

Methodology

1、LSTM模型搭建

LSTM模型有五个参数：输入维度、词向量维度、隐藏神经元数、隐藏层层数、输出维度，本实验中将词向量维度、隐层神经元数、隐藏层层数分别设置为30、512、2.损失函数选择交叉熵损失函数，初始学习率设置为0.001，使用Adam优化器。

2、数据预处理

读取16篇金庸小说，去除无意义的广告内容，使用jieba分词系统进行分词。每篇文章构建一个数据集，将中文文本转化成对应的索引方便输入到模型中。

3、评价指标

本实验需要对LSTM生成的文本进行定性和定量分析。定性方面，主要考察生成文本语义是否连贯，有无逻辑错误，与真实文本的相似程度。定量方面，使用困惑度（perplexity）这一指标进行衡量，困惑度越低说明模型性能越好。

Experimental Studies

1、实验结果

（1）结果1

提示文本：

是咱们同门表率何必亲自出马？让四师哥教训教训他也就是了。”鲁坤一听

模型生成文本：

是咱们同门表率何必亲自出马？让四师哥教训教训他也就是了。”鲁坤一听已明其意微笑道：“好四师弟咱们瞧你师兄不回来？”离座出去又要挥拳殴击。戚芳一把拉住叫道：“爹爹！”狄云大叫：“你们八个人再来打我有种的就一齐来。哪一个不来就是乌龟儿子狗杂种。”他急怒之下口不择言乱骂起来。万震山眉头一皱说道：“既是如此你们去领教狄师哥的高招。”五弟子卜垣极工心计昨晚见到狄云与万圭动手这乡下佬武功不弱这时情急拚命大师兄未必能胜如被他先赢得一仗纵然再有人将他打败也已折了万门的锐气同门中剑术以四师兄孙均为第一最好让孙均一上手便将他打败令他再也说嘴不得便道：“大师哥是咱们同门表率何必亲自出马？让四师哥教训教训他也就是了。”

实际文本：

是咱们同门表率何必亲自出马？让四师哥教训教训他也就是了。”鲁坤一听已明其意微笑道：“好四师弟咱们瞧你的了。”左手一挥七人一齐退开只剩孙均一人和狄云相对。孙均沉默寡言常常整天不说一句话上以能潜心向学剑法在八同门中最强。他见师兄弟推己出马当即长剑一立低头躬身这一招叫做“万国仰宗周衣冠拜冕旒”乃是极具礼的起手剑招。但当年戚长发向狄云说剑之时却将这招的名称说做“饭角让粽臭一官拜马猴”。意思是说：“我是好好的大米饭你是一只臭粽子外表上让你一下恭敬你一下我心里可在骂你！我是官你是猴子我拜你是官拜畜生。”狄云见他施出这一招心下更怒当下也是长剑一立低头躬身还了他一招“饭角让粽臭一官拜马猴”针锋相对毫不甘示弱。

（2）结果2

提示文本：

一柄明晃晃的匕首身边都是鲜血。窗子大开兀自摇晃戚长发却已不知去向。万圭

模型生成文本：

一柄明晃晃的匕首身边都是鲜血。窗子大开兀自摇晃戚长发却已不知去向。万圭怎样胜了狄云旁人见都没见到自然谈不上联手相攻了。沈城不过十五六岁年纪一副天真烂漫的样子谁都不信他会撒谎。万震山点了点头道：“原来如此。”戚长发气得满脸通红伸手一拍桌子喝道：“狄师兄瞧不起咱们要以一个斗八个难道咱们自己也瞧不起自己？”大弟子鲁坤道：“是众位师弟退开让我先领教领教狄师哥的高招。”

实际文本：

一柄明晃晃的匕首身边都是鲜血。窗子大开兀自摇晃戚长发却已不知去向。万圭哭叫：“爹爹！”扑到万震山身边。戚芳口中低声也叫：“爹爹！”身子颤抖握住了狄云的手。鲁坤叫道：“快快追凶手！”和周圻、孙均诸师弟纷纷跃出窗去大叫：“捉凶手捉凶手啊！”狄云见万门八弟子纷纷出去追赶师父这一下变故当真吓得他六神无主不知如何才好。戚芳又叫一声：“爹爹！”身子晃了两晃站立不定。

2、 结果分析

分析模型生成的两段文本，可以发现有一定的逻辑性但不够强，语句较通顺，与真实文本相比差距较大。

Conclusions

本次实验内容是使用LSTM模型生成文本，从结果上看，LSTM能够生成有一定语义逻辑的文本但与真实结果相比仍有所差距。