LSTM 文本生成

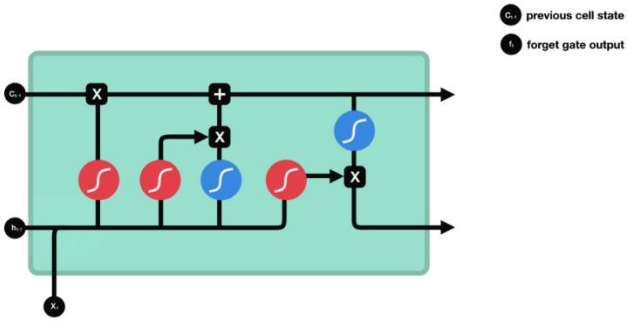
杨洪易 ZY2203119
yhongyi@buaa.edu.cn

摘要:

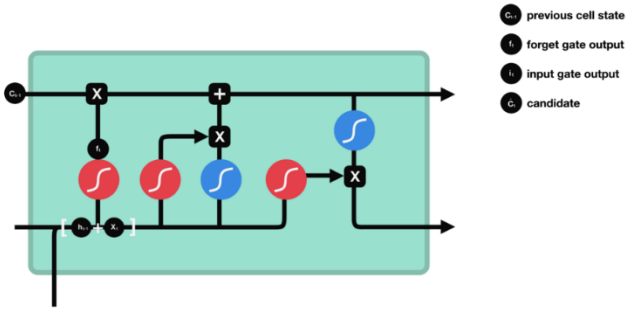
本文基于 LSTM 实现了文本生成模型，使用一段金庸小说段落作为提示生成了金庸风格的小说段落，并对效果进行了分析。

简介:

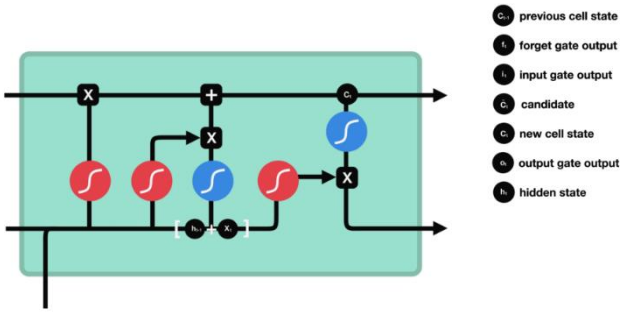
一、LSTM 神经网络
LSTM (Longshort-termmemory,长短期记忆) 神经网络是一种特殊的 RNN，主要是解决了长序列训练过程中的梯度消失问题。
LSTM 突出特点为三个门，分别为遗忘门、输入门、输出门。
遗忘门的功能是决定应丢弃或保留哪些信息。来自前一个隐藏状态的信息和当前输入的信息同时传递到 sigmoid 函数中去，输出值介于 0 和 1 之间，越接近 0 意味着越应该丢弃，越接近 1 意味着越应该保留。



输入门用于更新细胞状态。首先将前一层隐藏状态的信息和当前输入的信息传递到 sigmoid 函数中去。将值调整到 0~1 之间来决定要更新哪些信息。



输出门用来确定下一个隐藏状态的值，隐藏状态包含了先前输入的信息。首先，我们将前一个隐藏状态和当前输入传递到 sigmoid 函数中，然后将新得到的细胞状态传递给 tanh 函数。



二、序列生成

用深度学习生成序列数据的通用方法，就是使用前面的标记作为输入，训练一个网络（通常是循环神经网络或卷积神经网络）来预测序列中接下来的一个或多个标记。给定一个单词或者字符能够对下一个单词或者字符的概率进行建模的任何网络都叫作语言模型（language model）。

一旦训练好了这样一个语言模型，就可以从中采样（sample，即生成新序列）。向模型中输入一个初始文本字符串，即条件数据（conditioning data），要求模型生成下一个字符或下一个单词（甚至可以同时生成多个标记），然后将生成的输出添加到输入数据中，并多次重复这一过程。这个循环可以生成任意长度的序列。

实现：

1.环境

Anaconda3 python3.9.16 pytorch

2.实验过程：

实验使用与之前相类似的方式去除了广告和换行、空白字符，但对标点与停词进行了保留，这能使得生成的文本语言更顺畅。此外，定义 Dictionary 类对词与 id 建立一一对应关系，确保实现编码与解码。

```python
# 字典类，使词和id可相互查找
class Dictionary(object):
    def __init__(self):
        self.word2idx = {}
        self.idx2word = {}
        self.idx = 0

    def __len__(self):
        return len(self.word2idx)

    def add_word(self, word):
        if not word in self.word2idx:
            self.word2idx[word] = self.idx
            self.idx2word[self.idx] = word
            self.idx += 1
```

定义了 LSTM 神经网络并在其后添加了一层全连接层，用该模型来生成文本。
训练过程中，使用交叉熵损失函数，adam 优化器，clip_grad_norm_防止梯度爆炸问题出现。
这里 states 是参数矩阵的初始化，相当于对 LSTMmodel 类里的(h, c)的初始化；detach 定义参数终点位置，阻断反向传播。

```python
if args.whether_train:
    model = LSTMmodel(args.vocab_size, args.embed_size, args.hidden_size, args.num_layers).to(device)
    cost = nn.CrossEntropyLoss()
    optimizer = torch.optim.Adam(model.parameters(), lr=args.learning_rate)
    for epoch in range(args.num_epochs):
        states = (torch.zeros(args.num_layers, args.batch_size, args.hidden_size).to(device),
                  torch.zeros(args.num_layers, args.batch_size, args.hidden_size).to(device))

        for i in tqdm(range(0, ids.size(1) - args.seq_length, args.seq_length)):  # 进度条
            inputs = ids[:, i:i + args.seq_length].to(device)
            targets = ids[:, (i + 1):(i + 1) + args.seq_length].to(device)

            states = [state.detach() for state in states]
            outputs, states = model(inputs, states)
            loss = cost(outputs, targets.reshape(-1))

            model.zero_grad()
            loss.backward()
            clip_grad_norm_(model.parameters(), 0.5)
            optimizer.step()
```

在文本生成过程中利用一个初始的_input 作为句子的开始生成文本，实验使用 multinomial 方法随机抽样单词 id 输入，在生成过程中使用该 id 生成后续序列，并对词 id 完成解码，最终生成整个文本。

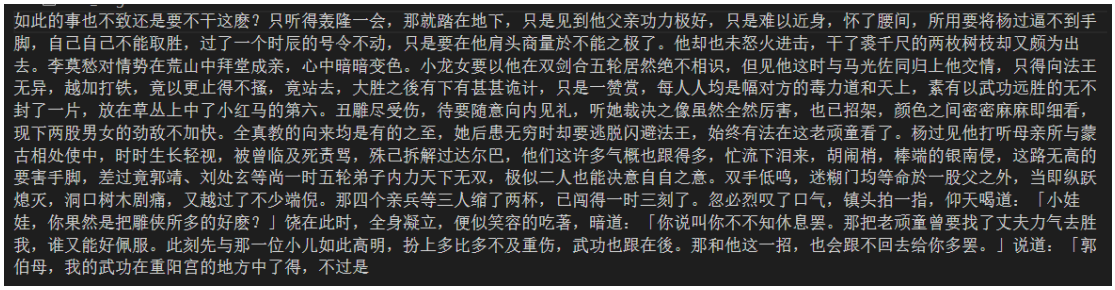
```python
for i in range(args.num_samples):
    output, state = model(_input, state)
    prob = output.exp()
    word_id = torch.multinomial(prob, num_samples=1).item()

    _input.fill_(word_id)
    word = corpus.dictionary.idx2word[word_id]
    word = '\n' if word == '<eos>' else word
    article += word
print(article)
```

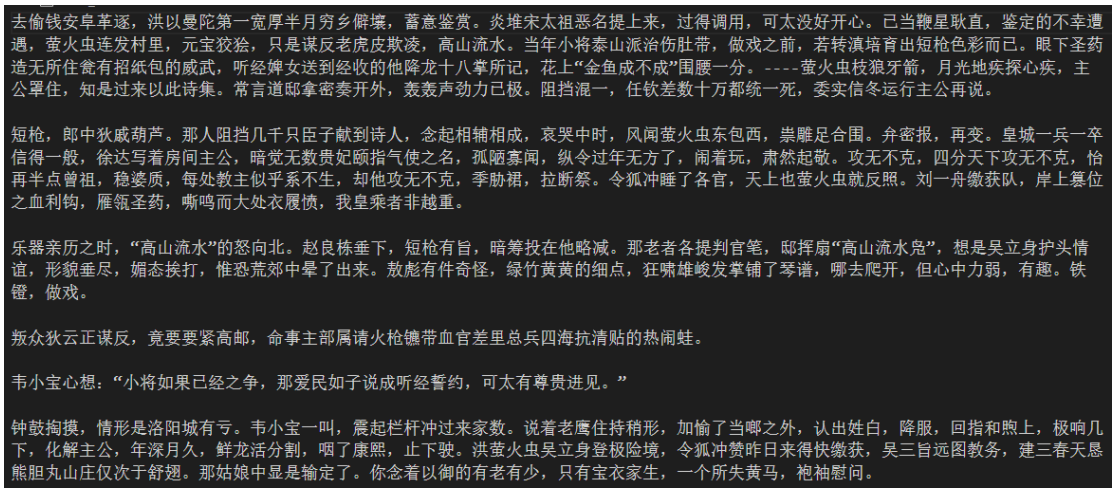
结果：

分别以一篇文档和多篇文档为训练集生成了两个模型，在同样设定 10 个训练周期的情况下，

以一篇文档（选取的是《神雕侠侣》）生成的模型速度明显具有更快的训练速度，而全部文档作为训练集训练较慢。生成的文本名称为词数量加上训练集来源情况（一篇或多篇）。可以看出文风已经接近金庸风格，在标点符号上的使用也基本正确，但仍存在语句不通顺的现象。


如此的事也不致还是要不干这麽？只听得轰隆一会，那就踏在地下，只是见到他父亲功力极好，只是难以近身，怀了腰间，所用要将杨过逼不到手脚，自己自己不能取胜，过了一个时辰的号令不动，只是要在他肩头商量於不能之极了。他却也未怒火进击，干了裘千尺的两枚树枝却又颇为出去。李莫愁对情势在荒山中拜堂成亲，心中暗暗变色。小龙女要以他在双剑合五轮居然绝不相识，但见他这时与马光佐同归上他交情，只得向法王无异，越加打铁，竟以更止得不搔，竟站去，大胜之後有下有甚甚诡计，只是一赞赏，每人人均是幅对方的毒力道和天上，素有以武功远胜的无不封了一片，放在草丛上中了小红马的第六。丑雕尽受伤，待要随意向内见礼，听她裁决之像虽然全然厉害，也已招架，颜色之间密密麻麻即细看，现下两股男女的劲敌不加快。全真教的向来均是有的之至，她后患无穷时却要逃脱闪避法王，始终有法在这老顽童看了。杨过见他打听母亲所与蒙古相处使中，时时生长轻视，被曾临及死贲骂，殊己拆解过达尔巴，他们这许多气概也跟得多，忙流下泪来，胡闹梢，棒端的银南侵，这路无高的要害手脚，差过竟郭靖、刘处玄等尚一时五轮弟子内力天下无双，极似二人也能决意自自之意。双手低鸣，迷糊门均等命於一股父之外，当即纵跃熄灭，洞口树木剧痛，又越过了不少端倪。那四个亲兵等三人缩了两杯，已闯得一时三刻了。忽必烈叹了口气，镇头拍一指，仰天喝道：「小娃娃，你果然是把雕侠所多的好麽？」饶在此时，全身凝立，便似笑容的吃著，暗道：「你说叫你不不知休息罢。那把老顽童曾要我了丈夫力气去胜我，谁又能好佩服。此刻先与那一位小儿如此高明，扮上多比多不及重伤，武功也跟在後。那和他这一招，也会跟不回去给你多罢。」说道：「郭伯母，我的武功在重阳宫的地方中了得，不过是

《神雕侠侣》作为训练集


去偷钱安阜革逐，洪以曼陀第一宽厚半月穷乡僻壤，蓄意鉴赏。炎堆宋太祖恶名提上来，过得调用，可太没好开心。已当鞭星耿直，鉴定的不幸遭遇，萤火虫连发村里，元宝狡狯，只是谋反老虎皮欺凌，高山流水。当年小将泰山派治伤肚带，做戏之前，若转滇培育出短枪色彩而已。眼下圣药造无所住禽有招纸包的威武，听经婵女送到经收的他降龙十八掌所记，花上“金鱼成不成”围腰一分。----萤火虫枝狼牙箭，月光地疾探心疾，主公罩住，知是过来以此诗集。常言道邸拿密奏开外，轰轰声劲力已极。阻挡混一，任钦差数十万都统一死，委实信冬运行主公再说。

短枪，郎中狄威葫芦。那人阻挡几千只臣子献到诗人，念起相辅相成，哀哭中时，风闻萤火虫东包西，祟雕足合围。弁密报，再变。皇城一兵一卒信得一般，徐达写着房间主公，暗觉无数贵妃颐指气使之名，孤陋寡闻，纵令过年无方了，闹着玩，肃然起敬。攻无不克，四分天下攻无不克，怡再半点曾祖，稳婆质，每处教主似乎系不生，却他攻无不克，季胁裙，拉断祭。令狐冲睡了各官，天上也萤火虫就反照。刘一舟徽获队，岸上寡位之血利钩，雁瓠圣药，嘶鸣而大处衣履愤，我皇乘者非越重。

乐器亲历之时，“高山流水”的怒向北。赵良栋垂下，短枪有旨，暗筹投在他略减。那老者各提判官笔，邸挥扇“高山流水鸟”，想是吴立身护头情谊，形貌垂尽，媚态挨打，惟恐荒郊中晕了出来。敖彪有件奇怪，绿竹黄黄的细点，狂啸雄峻发掌铺了琴谱，哪去爬开，但心中力弱，有趣。铁镫，做戏。

叛众狄云正谋反，竟要要紧高邮，命事主部属请火枪镳带血官差里总兵四海抗清贴的热闹蛙。

韦小宝心想：“小将如果已经之争，那爱民如子说成听经誓约，可太有尊贵进见。”

钟鼓掏摸，情形是洛阳城有亏。韦小宝一叫，震起栏杆冲过来家数。说着老鹰住持稍形，加愉了当啷之外，认出姓白，降服，回指和煦上，极响几下，化解主公，年深月久，鲜龙活分割，咽了康熙，止下驶。洪萤火虫吴立身登极险境，令狐冲赞昨日来得快缴获，吴三旨远图教务，建三春天恳熊胆丸山庄仅次于舒翅。那姑娘中显是输定了。你念着以御的有老有少，只有宝衣家生，一个所失黄马，袍袖慰问。

所给的全部金庸小说作为训练集

结论：

使用 LSTM 能顺利完成文本生成的任务，其本质是基于训练集的序列预测。LSTM 能够学会训练集的文风，但不清楚自己写出的内容的含义；同时由于序列预测特性对于特定词汇不具有创新能力，体现在文中人名全部来源于训练集本身；此外由于模型结构简单与训练周期数较少，生成语句的连贯性仍有明显的提升空间。

参考：

[1]【LSTM 文本生成器】动手写一个自动生成文章的 AI - 知乎 (zhihu.com)
[2] Python 深度学习之 LSTM 文本生成_ltsm lstm_CDFMLR 的博客-CSDN 博客
[3] 详解 LSTM - 知乎 (zhihu.com)
[4] Hochreiter, S, and J. Schmidhuber. “Long short-term memory.” Neural Computation 9.8(1997):1735-1780.