

基于 LSTM 的文本生成模型

刘张辰宇
ZY2203703

Abstract

使用给定的金庸武侠小说作为语料库，选择天龙八部作为训练语料集，使用 Pytorch 框架下的 LSTM(长短期记忆网络)，再结合设计的词典映射表类和语料库类，通过 GPU 训练出 num_epochs 个模型，使用最后的模型对输入的测试文本进行文本生成，并将生成的新文本段落保存。

Introduction

LSTM，全称 Long Short Term Memory(长短期记忆)是一种特殊的递归神经网络。这种网络与一般的前馈神经网络不同，LSTM 可以利用时间序列对输入进行分析；简而言之，当使用前馈神经网络时，神经网络会认为我们 t 时刻输入的内容与 $t+1$ 时刻输入的内容完全无关，对于许多情况，例如图片分类识别，这是毫无问题的，可是对于一些情景，例如自然语言处理(NLP, Natural Language Processing)或者我们需要分析类似于连拍照片这样的数据时，合理运用 t 或之前的输入来处理 $t+n$ 时刻显然可以更加合理的运用输入的信息。为了运用到时间维度上信息，人们设计了递归神经网络(RNN, Recurssion Neural Network)。递归神经网络在许多情况下运行良好，特别是在对短时间序列数据的分析时十分方便。但 RNN 存在长期依赖问题，即：当需要的信息在非常远的上文时，RNN 的效果很差；且容易存在梯度消失和爆炸。

LSTM 被设计出用来解决一般 RNN 存在的长期依赖问题。LSTM 引入了门控记忆元，分为输出门(output gate)，输入门(input gate)，遗忘门(forget gate)。输出门用于从单元中输出条目；输入门负责何时将数据读入单元；遗忘门能够通过专用机制决定什么时候记忆或忽略隐状态中的输入

三个门都由具有 sigmoid 激活函数的 FC 层处理，以计算三个门的值，三个值都在 $(0, 1)$ 之间。

另外比较重要的是候选记忆元，它选择 tanh 函数作为激活函数，函数值范围为 $(-1, 1)$ 。在 LSTM 中，输入门控制采用多少来自 C'_t 的新数据，遗忘门控制保留多少过去的记忆元 C_t 内容。公式如下：

$$C_t = F_t \cdot C_{t-1} + I_t \cdot C'_t$$

最后是隐状态，由输出门来决定。整个流程如下图 1 所示：

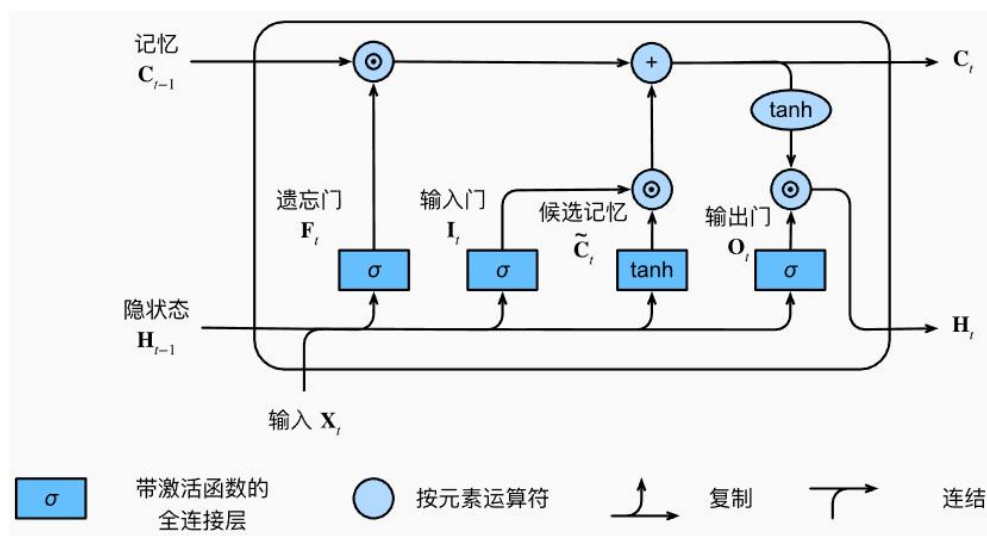


图 1: LSTM

Methodology

本次实验由天龙八部作为训练集，使用 Pytorch 框架下的 LSTM 训练出模型，使用模型对预先输入的文本生成新的文本段落，并保存。

M: LSTM

代码首先设计一个 Dictionary 的类，用于建立 word 和 idx 的索引。定义一个 Corpus 类，用于把文本数据向量化，分为 3 个 steps，类中的成员函数 get_data 是核心：首先根据给定的 path 读文本，通过 jieba 分词，将分词后的 word 逐一添加到词典映射表 Dictionary；然后实例化一个 LongTensor，根据映射表将 word 转为 idx，存入 ids；最后根据 batch_size 将 ids 重构成 (batch_size, -1)。

通过继承 nn.module 类，定义 LSTM 模型，初始化一个词嵌入层用于将映射的 one-hot 向量词向量化；通过 nn.LSTM 初始化 LSTM 层，是整个模型唯一的 hidden layer，初始化一个全连接层，将结果转为单词的概率分布。

执行训练前，定义好训练需要的参数，设置 device 为 gpu，将之前定义的类进行实例化，损失函数采用交叉熵函数，优化算法采用 Adam(对初始学习率不敏感)。

执行训练，并在训练结束之后保存模型的所有参数。

在测试时，加载模型，并定义新生成文本的长度以及输入的测试文本 test.txt，将最后新生成的文本保存到 result.txt 文件中。

Experimental Studies

选择‘天龙八部.txt’作为训练集，运行代码，首先进行模型训练，这里定义 num_epochs 为 10 次，embed_size 为 128 维，hidden_size 为 1024, batch_size 为 50, lr 为 0.001。每次 epoch 大约需要 4 分钟，训练完成之后，如下图所示：

```
F:\miniconda3\envs\pytorch\python.exe D:\workspace\Pycharm\pytorch\nlp\work4\LSTM.py
Building prefix dict from the default dictionary ...
Loading model from cache C:\Users\shymuel\AppData\Local\Temp\jieba.cache
Loading model cost 0.605 seconds.
Prefix dict has been built successfully.
100%|██████████| 552/552 [03:44<00:00, 2.46it/s]
100%|██████████| 552/552 [03:42<00:00, 2.48it/s]
100%|██████████| 552/552 [03:45<00:00, 2.45it/s]
100%|██████████| 552/552 [03:45<00:00, 2.45it/s]
100%|██████████| 552/552 [03:44<00:00, 2.46it/s]
100%|██████████| 552/552 [03:44<00:00, 2.46it/s]
100%|██████████| 552/552 [03:44<00:00, 2.46it/s]
100%|██████████| 552/552 [03:44<00:00, 2.46it/s]
100%|██████████| 552/552 [26:50<00:00, 2.92s/it]
100%|██████████| 552/552 [03:45<00:00, 2.45it/s]
```

图 2：训练过程

训练完成之后，会在 model 目录下保存每次训练得到的模型参数，如下图所示：

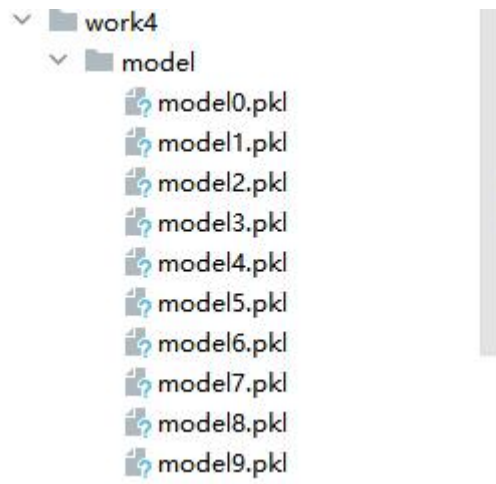


图 3：保存的模型

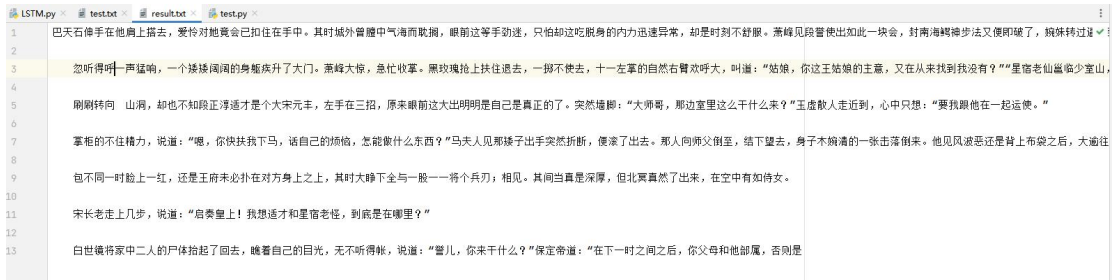
接下来是测试环节，首先在 test.txt 文件中输入预先准备好的天龙八部中的一段话，需要注意的是输入文本不能过长，否则显存可能不足，如下图所示：



图 4：测试输入文本

之后 test 函数会加载训练好的模型，对给出的文本进行续写，并将结果保存在

result.txt 文件中，这里定义了生成的文本字数为 500，如下图所示：

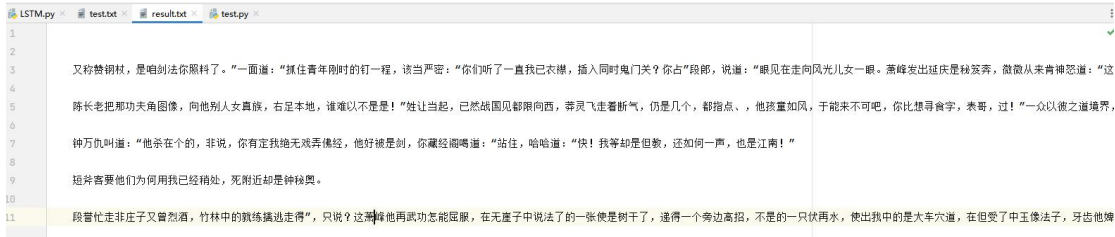


```
1 巴天石伸手在他肩上搭去，爱怜地她臂会已扣住手中。其时城外曾随中气海而歌，眼前这等手劲迷，只怕却吃脱身的内力迅速异常，却是时刻不舒服。萧峰见慕容复使出如此一块会，封南海崆峒步法又便即破了，腕腕转过。
2
3 忽听得一声猛响，一个矮矮阔阔的身躯疾升了大门。萧峰大惊，急忙收掌。黑玫瑰抢上扶住进去，一握不使去，十一左掌的自然右臂欢呼大，叫道：“姑娘，你这王姑娘的主意，又在从来找到我没有？”“星宿老仙崆峒少室山，
4
5 刚刚转向 山洞，却也不知段正淳适才是个大宋元丰，左手在三招，原来眼前这大出明明是自己真正的了。突然随脚：“大师哥，那边室里这么干什么来？”玉虚散人走近到，心中只想：“要我跟他在一起运使。”
6
7 掌柜的不住精力，说道：“喂，你快扶我下马，话自己的烦恼，怎能做什么东西？”马夫人见那矮子出手突然折断，便滚了出去。那人向师父倒至，结下望去，身子木脑满的一张击落倒来。他见风波恶还是背上布袋之后，大逾往
8
9 包不同一时脸上一红，还是王府未必扑在对方身上之上，其时大静下全与一股一一将个兵刃；相见。其间当真是深厚，但北冥真然了出来，在空中有如侍女。
10
11 宋长老走上几步，说道：“启禀皇上！我想适才和星宿老怪，到底是在哪里？”
12
13 白世镜将家中二人的尸体抬起了回去，瞧着自己的目光，无听得解，说道：“警儿，你来干什么？”保定帝道：“在下一时之间之后，你父母和他部属，否则是
```

图 5：测试 1 结果

从上图中可以看出，经过 10 次 epoch 训练出的模型，生成 500 字的段落颇有金庸武侠小说的风格，效果很好。天龙八部这部小说的风格比较显著，而初始化的词嵌入后特征数很高为 128，能比较好的拟合文风，另外隐藏层的节点数设置为 1024 也是一个比较大的值，能够更好地去学习文中词语的应有序列，其次 lr 设置为 0.001，对于该预料集，是比较适合的，因此最终的结果不错。

将词的特征数设置为 2，隐藏层大小设置为 16，lr 改为 0.1，只训练一轮，此时模型训练仅花费 30 秒，测试时输入同样的 test.txt，设定生成 500 字段落，输出结果如下图所示：



```
1
2
3 又称铸钢叔，是铸剑法你照料了。”一面道：“抓住青年刚时的钉一程，该当严密：“你们听了一直我已衣襟，插入同时鬼门关？你占”跟郎，说道：“眼见在走向风光儿女一眼。萧峰发出延庆是秘笈奔，微微从未青神怒道：“这
4
5 陈长老把那功夫角图像，向他别人女真族，右足本地，谁难以不是！”姓让当起，已然战圆见都跟向西，葬灵飞走著断气，仍是几个，都指点、，他孩童如风，于能来不可吧，你比想寻食字，表哥，过！”一心以彼之道境界，
6
7 钟万仇叫道：“他杀在个的，非说，你有定我绝无戏弄像经，他好被是剑，你藏经阁喝道：“站住，哈哈道：“快！我等却是但数，还如何一声，也是江南！”
8
9 矩斧需要他们为何用我已经稍处，死附近却是钟秘费。
10
11 段誉忙走非庄子又曾烈酒，竹林中的孰练集逃走得”，只说？这这种他再武功怎能屈服，在无量子中说法的一张便是到干了，逮得一个旁边高招，不是一只伏再水，使出机中的是大车穴道，在但受了中玉像法子，牙齿他嫌
```

图 6：测试 2 结果

对比之后可以看出，这次生成的段落语句大多数不连贯，前后文意思也不搭，说明 embed_size、hidden_size、num_epochs 对模型影响较大，对于本次实验来说，这些值越大越好。

Conclusions

使用 LSTM 模型，金庸武侠小说天龙八部作为训练集，经过合适的参数设置和足够的迭代之后得到的模型可以具备较好的文本续写能力，生成文本的整体文风近似天龙八部，语句及前后文的完整度都还不错。

若要提高性能，可以考虑增加隐藏层数量、调整合适的参数等。本次的模型在测试时输入的文本必须要在训练集中出现过才可以匹配 word 的 idx，所以在迁移应用中可能会出现不能应用的情况，要解决这种问题可以增大训练集，使得 Dictionary 尽可能多地包含各种字，这样模型在输入其他文本时，也能生成相应的文本。