基于 LSTM 实现文本生成模型

彭海琅 ZY2203322

plunck@buaa.edu.cn

摘要

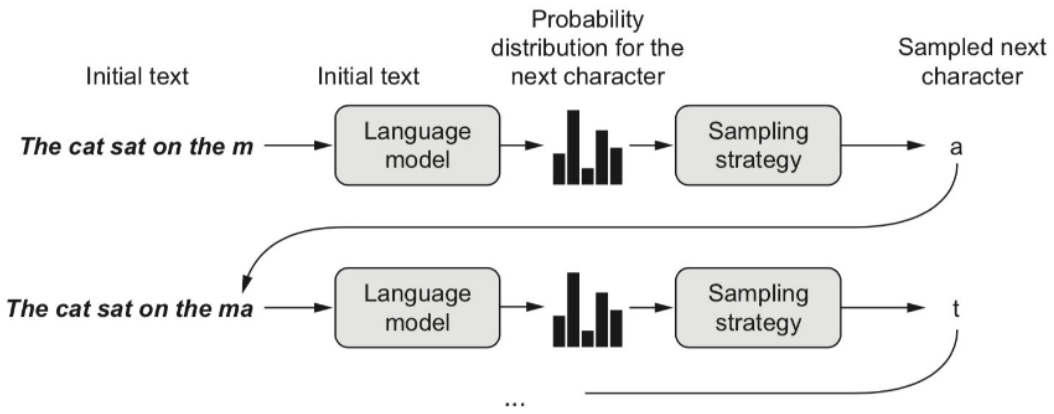
基于深度学习生成序列的通用方法，就是训练一个网络(一般用 RNN 或 CNN)，输入前面的 Token，预测序列中接下来的 Token。长短期记忆网络（LSTM，Long Short-Term Memory）是一种时间循环神经网络，是为了解决一般的 RNN（循环神经网络）存在的长期依赖问题而专门设计出来的。LSTM 的表现通常比时间递归神经网络及隐马尔科夫模型（HMM）更好，比如用在不分段连续手写识别上。本文实现了以金庸小说原文作为训练文本，并根据输入生成新的文本。

导论

一、序列生成之文本生成

基于深度学习生成序列的通用方法，就是训练一个网络(一般用 RNN 或 CNN)，输入前面的 Token，预测序列中接下来的 Token。给定前序的 token，能够对下一个 token 的概率进行建模的网络叫做语言模型。语言模型能够捕捉到语言的统计结构，当训练好一个语言模型后，输入初始的文本字符串(称为条件数据)，从语言模型中采样，就可以生成新 token，把新的 token 加入条件数据中，再次输入，重复这个过程就可以生成任意长度的序列。

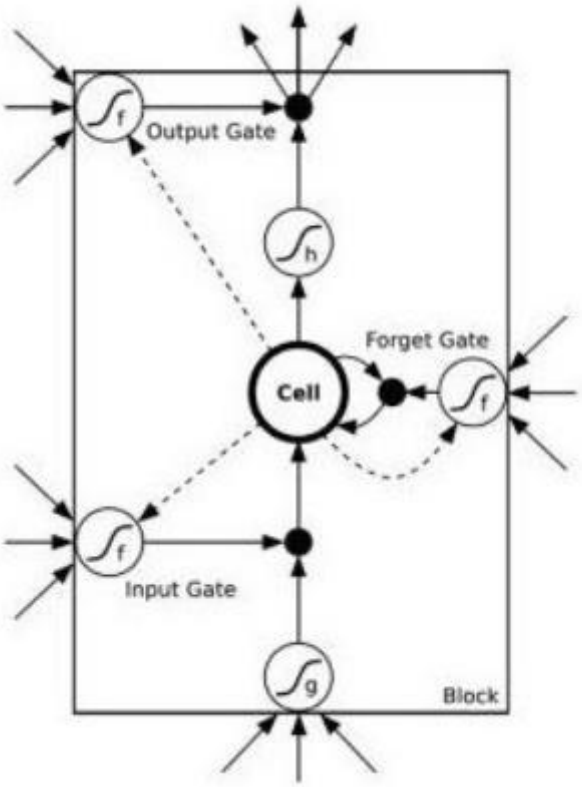
在自然语言处理(NLP)领域，大多对话机器人的对话形成都会采用基于语料库和深度神经网络生成模型进行回答和交流。很多企业成功落地了许多产品，例如，微软小冰、Siri、谷歌翻译、淘宝智能客服系统等。文本生成的商业价值不断提升，用户的要求也不断提高，因此文本生成的研究意义重大。此外，在一些自动文本摘要和文本简化的地方，也会采用神经网络生成一段可读性强而且易于常人理解的文本。目前主流的文本生成的神经网络大多是基于循环神经网络的编码器-解码器框架，通过循环神经网络学习到序列的前后关系，能够生成一些较为通顺的文本。



二、长短期记忆网络

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。它是一种含有 LSTM 区块（blocks）或其他的一种类神经网络，文献或其他资料中 LSTM 区块可能被描述成智能网络单元，因为它可以记忆不定时间长度的数值，区块中有一个 gate 能够决定 input 是否重要到能被记住及能不能被输出 output。

LSTM 的表现通常比时间递归神经网络及隐马尔科夫模型（HMM）更好，比如用在不分段连续手写识别上。



实验内容

一、问题描述

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。

二、实现步骤

1、读取文本，建立字典信息，把全文信息转换成带编号的长向量。

2、构建 LSTM 网络，第一层用 Embedding 层，初始化一个词嵌入层，用来将映射的 one-hot 向量词向量化，第二层用 LSTM 层，第三层为 Dense 层，用来把神经网络的运算结果转化为单词的概率分布。。

3、进行迭代训练，并比较不同的输出结果之间的差异。

实验结果

词典大小：3047
第一个词：
越

epoch:0/100 process:5/6 loss:6.664
暗暗大量出现一跃跟随日日夜夜使臣不许圆小不忍则乱大谋还要羊。一阵已口得瞎，剑术割去，可竹棒不是守招这又向财力

epoch:26/100 process:5/6 loss:4.665
尽聚在一起国剑士计策，半截使开，令越起来，回禀以多为骇人听闻，只须提起吩咐青衣剑士缓缓真是并肩手腕和这八人。”范蠡同时

epoch:48/100 process:5/6 loss:2.470
勾践召见伍子胥蹲下匣中。炽热的血两名左手。来众人道：“姑娘，我想拔他挡住，这样大王并肩两三年的转动

epoch:80/100 process:5/6 loss:0.656
学剑之士报仇，八人轻响？”突然一见王者听得越，定得吴国剑士，只见他如何在核心。回去、这

epoch:83/100 process:5/6 loss:0.571
学剑之士，当守说道：“啊哟”范蠡道：“你真是嫌屋子的草地，风胡子”铸剑。”他抬头之下

epoch:90/100 process:5/6 loss:0.406
城内城外，荒山野岭中去找寻，在兄长温柔去一口，在真是孙武子兵法，剑戟孙武子王僚。薛烛之说道

epoch:92/100 process:5/6 loss:0.362
薛烛召见薛烛，又两名青衣剑士以二对的身子被空中，又杀了她，要要得此人给小人师兄，他铸剑

epoch:99/100 process:5/6 loss:0.247
米一阵摆脱”小人等手指，然则我士夷光，从逼这句话之中，已知访。”勾践道：“

从中可以看出，随着训练迭代次数的增加，生成的文本变得越来越有逻辑，也就是说更接近人话。

结论

本文实现了以金庸小说原文作为训练文本，并根据输入生成新的文本。因为训练次数少和模型比较简单，生成的结果并不完美，然而 LSTM 作为一种优秀的循环神经网络 cell 是非常有潜力的。