深度学习与自然语言处理第四次报告

郭子龙

1216199336@qq.com

摘要

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。

1. 介绍

人类的感知模式、语言和艺术作品都具有统计结构。学习这种结构是深度学习算法所擅长的。机器学习模型能够对图像、音乐和故事的统计潜在空间（latent space）进行学习，然后从这个空间中采样（sample），创造出与模型在训练数据中所见到的艺术作品具有相似特征的新作品。

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，所有的RNN都具有一种重复神经网络模块的链式形式。在标准RNN中，这个重复的结构模块只有一个非常简单的结构，例如一个tanh层。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。2009年，用LSTM构建的人工神经网络模型赢得过ICDAR手写识别比赛冠军。LSTM还普遍用于自主语音识别，2013年运用TIMIT自然演讲数据库达成17.7%错误率的纪录。作为非线性模型，LSTM可作为复杂的非线性单元用于构造更大型深度神经网络。

2. 方法

M1: 数据读取与预处理模块

第一部分对文本进行读取，按照所给目录对小说进行读取和存储，之后对小说进行预处理，首先删去原文中的标点符号和广告词。随后依据停词语料库删去文本中相关停词。

M2: 生成序列数据

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。

M3: 生成文本的采样策略

采样方式分为贪婪采样（greedy sampling）与随机采样。其实前一种可以看作后一种的特例。随机就是根据sofmax输出的概率进行采样，如果预测的一个词概率是10%，那么它被采样的概率也是10%。为了控制采样随机性的大小，我们引入softmax温度这个参数，对概率重新进行加权并计算。温度越高随机性越大，温度取0，就是贪婪采样。

3. 实验结果

输入一：

到张朝唐十二岁那一年，福建有一名士人屡试不第，弃儒经商，随着乡人来到渤泥国。这人不善经营，本钱蚀得干干净净，无颜回乡，

输出一：

心中怦怦乱跳，忽然不答，心下又惊又喜，低声道：“海老公这一味好，下恨到底，不杀的。”苏荃一笑，过不四拜，自己又如何救礼，恸然，当时暗一球扇散走，怎算得已啊擒马啊？”太后又道：“认错，从此大家不提亲，他立刻一面对付你的材，什么也是说话外。到我会没去，庄家不行，后会就收我，也得在关外陪到那里的去等要蒙成一本书来，居然还不迟，骗拳吴立子要打败我的啊，果不是高大人。咱们回去？为了我心好不快，便冲着你。”那老者道：“完，好不把穴道付。你骗得事就真闲见，我不敢瞧瞧我，就此同时小，要放在桌子上埋赐了，说不定是形势？在中原来自然是平平西王，要骗人大量便有了大封怪事，那单强贪如的仙福，叫做英雄都带来学，不过戴上五行江就程，你要长喝运气，哈哈，. . . 我. . . 你放心倒

输入二：

众人见这道人身上道袍葛中夹丝，灿烂华贵，道冠上镶着一块晶莹白玉，光华四射，背负长剑，飘飘然有出尘之概，约莫四五十岁年纪，一身清气，

输出二：

一翻子，轻轻和荆棘相沅棚成？清水知道之识对十三天拜波，无法爪倒。那派人家不醒，飞起来送到他腿上。济尔祝郁闷也不问，只见他跳动板臂，连膝跪倒，身边对准了门下口，两点同乐术，那是的十余人是出口鲜血。只见一丛手弯弯曲曲，在小抚后和各口同时跃得靴里，鲜血直流。只听丁典道：“姓胡的这小子的功夫，是，那也不怕师叔算好端了，这男子可也学得，快快……好。”那老乞丐从怀中取出一银两面的郡女走来，躺在桌上，双儿说道：“什么地方？一生之中，你大小骗腋，可得给太老婆的、从不过，也不用，只得但天地会的，难道你此人只知没有再骑，不过二十三个空掌门人大会之物，伙错得来干冒犯了折了咱们下，可和只见出他的身上，那是搞的，也没想到世印消给那一株半点也不死，是拿过个人，先剌杀为一件双谨为王面，这可不做龙脉之事，教主疯人已忘了为何用就，我世甚会做老伯子请，别惹人啦。你从此就瞒不民已毕，不过随师父的怪事就给你弄瞎改嘴。老嘴走上官，别让你

4.总结

此次模型文本生成实验，对算法、数据库、显存设备等要求较高，能够实现基本功能，但输出结果不尽如人意，在后续学习过程中，期望改进算法，利用更多的数据库训练，以期得出更好的输出。炸显存问题较为严重。