利用 LSTM 进行文本生成

ZY2203810

吴金旺

Abstract

本文简单介绍了 LSTM 模型，并用它对金庸和古龙的小说进行文本生成。

Introduction

循环神经网络（Recurrent Neural Network，RNN）是一种用于处理序列数据的神经网络。相比一般的神经网络来说，它能够处理序列变化的数据。

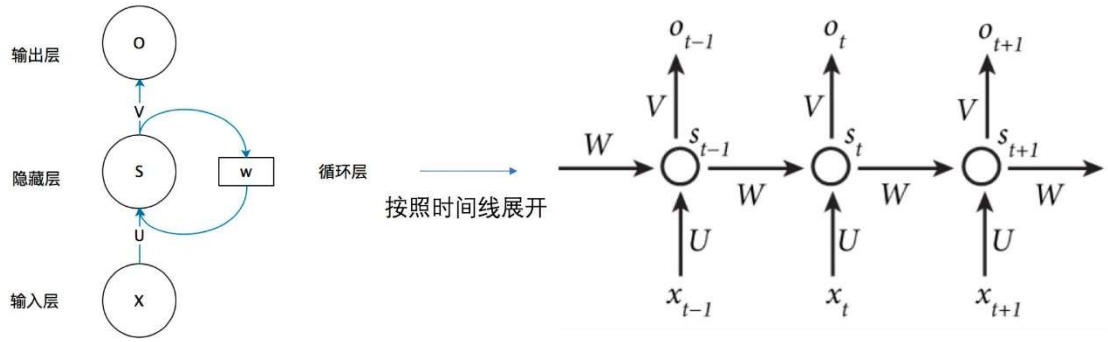

图 1 RNN 网络示意图

一个典型的 RNN 如图 1 所示，每一时刻的输出取决于当前时刻的输入与之前的隐藏层输出，该时刻的隐藏层输出又可以继续用于下一时刻的预测。通常，两个状态之间仅有一个简单的线性层和一个激活函数。

RNN 结构简单，能较好地处理较短的序列的预测问题，但是随着序列长度的增加，序列早期的内容会经过多次的计算才能传递到序列的末端，这导致其信息会被遗忘，此外还会有梯度消失/爆炸的问题。虽然在序列预测中，一般离当前时刻较近的内容更有用，但是也存在一些离当前时刻较远的关键信息。

长短时记忆网络（Long Short Term Memory，LSTM）是一种特殊的 RNN，主要是为了解决长序列训练过程中的梯度消失和梯度爆炸问题。相比普通的 RNN，LSTM 能够在更长的序列中有更好的表现。

一个典型的 LSTM 如下图所示，前一时刻的输出h_{i-1}和“记忆”C_{i-1}会输入到这一时刻的网络中。然后依次经过三个门：

1) 遗忘门：每一次得到一个新的输入，LSTM 会先根据新的输入和上一时刻的输出决定遗忘掉之前的哪些记忆，有$f_t = \sigma\big(W_f \cdot [h_{t-1}, x_t] + b_f\big)$。$f_t$决定了要忘记哪些内容。它会与记忆$C_{i-1}$相乘，保留其中的有效信息，遗忘无效信息。

2) 记忆门：记忆门决定当前时刻的输入有多少要被记住，有$C'_t = tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$，$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$。两者会先做一个乘法，这就是当前时刻实际要记住的内容，它会与已经通过遗忘门的记忆做加法，得到当前时刻更新后的记忆C_t，即$C_t = f_t \times C_{t-1} + i_t \times C'_t$。

3) 输出门：输出门决定当前时刻的输出，它根据更新后的记忆和当前时刻的输入决定输出，有$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$，$h_t = o_t \times tanh(C_h)$。

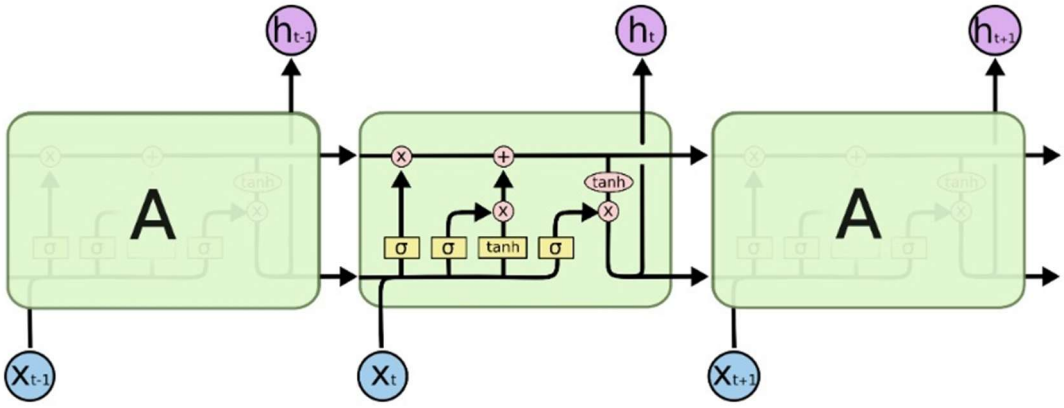

图 2 LSTM 结构图

经过三个门后，LSTM 记住了有效的信息，遗忘了无效的信息，这使得它可以在较长的序列预测中，始终保持的长时记忆，而当前的输入又可以认为是短时记忆，这也是 LSTM 名字的由来，它也较好的符合了人的记忆特点。

Methodology

LSTM 可以用于文本的生成，因为文本本身就是一个序列，且一般非常长，每句话中也会有较重要的动词、名词和不那么重要的介词等，那么普通的 RNN 就会受遗忘问题的影响，而 LSTM 则相对没有这个问题。

在利用 LSTM 进行文本生成时，首先要给定一段初始的文本，假设长度为$t-1$，这段文本可以用于计算$t-1$时刻状态C_{t-1}，以及输出h_{t-1}，然后正式开始预测。

一般我们利用自回归的 LSTM 进行生成任务。即在预测时，将上一时刻的输出作为该时刻的输入，这样可以通过 LSTM 网络得到该时刻的输出，此输出作为下一时刻的输入，继续计算，最终生成整个文本序列。具体来说，有：

$$x_t = h_{t-1}$$

$$h_t, C_t = LSTM(C_{t-1}, x_t)$$

这样仅需提供初始的一小段文本，就可以不停的往后生成文本，直到输出达到设定的上限，或者输出结束标志时停止。

在训练时，有两种训练方法。由于训练时文本全部可见，第一种方法是每一时刻的输入都是真实的文本，然后输出仅仅用于比较与真实文本的区别，并用于反向传播。另一种方法是在一定长度的序列范围内，输出直接作为下一时刻的输入，直到生成整个序列，然后将它整个与真实的文本序列做比较。相比较而言，后者更接近真实的生成任务，但难度较大，需要大量文本和性能较好的模型；前者难度更低，少量文本即可训练。本次实验主要利用第一种方法进行训练。

Experimental Studies

A. 实验设定和数据集

我们利用一个隐藏层为 1024（即记忆的维度为 1024）的单层 LSTM 进行训练。我们的字典里是一个个的标点符号和汉字，利用一个 embedding 层将它们转换为 256 维的向量，从而进行后续的计算。

语料库是已经给的金庸小说集和网上查找得到的古龙小说集，一共约 80 篇小说。我们首先对它进行过滤，去除英文，然后统计语料库里的所有汉字和标点符号，制作成一个字典，用于上文提到的文字向量化。

我们在每一个自然段后面加上一个结束符<eos>，然后将所有文本拼接在一起，按照一定的序列长度进行分割，这里我们的序列长度设为 256，作为训练集的输入文本。而其输出文本，或者说是预测结果，则是从输入文本的第二个 token 开始，到它结束的后一个 token 结束的另一个长度为 256 的序列，也就是与它相差一个 token 的序列。

初始状态设为 0，然后每一个序列的最终状态作为下一个序列的输入状态，这样将所有的序列遍历，就实现了训练过程。

测试阶段，将随机从数据集中抽取一定的文本进行测试。虽然存在测试集泄露的问题，但文本生成效果实在太差，即使是这样也无法完整的生成后续文本。计算文本相似度的方法十分简单，我们将一句话的每个词都用 embedding 层转换为向量，然后计算向量的平均值，作为这一句话的向量，然后计算向量之间的余弦相似度，定量的判断文本生成的质量。

B.运行结果

利用金庸和古龙小说集进行训练后，测试选取了其中的 200 个段落，目标是生成它们后面一段的内容。

经测试，平均的文本得分为 0.509，虽然不高，但也说明至少具有一定的相关性，不过这一得分其实没有太大意义。实际考察生成的文本，发现质量很差，如：

原文：他初时左掌拍出，知道周芷若武功与自己已相差不远，大是强敌，丝毫不敢怠忽，加之单掌迎双掌，这一掌乃是出了十成力，劲力刚向外吐，便即察觉对方力尽，急忙硬生生的收回，他明知这是犯了武学的大忌，等于以十成掌力回击自身，何况在这间不容发之际突然回收，用力更是奇猛，但他于自己内劲收发由心，这股强力回撞，最多一时气窒，决无大碍。不料他掌力刚回，突觉对方掌力犹似洪水决堤、势不可当的猛冲过来。张无忌大吃一惊，知道已中暗算，胸口砰的一声，已被周芷若双掌击中。那是他自己的掌力再加上周芷若的掌力，并世两大高手合击之下，他护体的九阳神功虽然浑厚，却也抵挡不住。何况周芷若的掌力乃乘隙而进，正当他旧力已尽、新力未生之时。这门功夫却是峨嵋派嫡传，当年灭绝师太便曾以此法击得他喷血倒地。只不过当年他是全然不知抵御，这次却是一念之仁、受欺中计。当下不由自主的身向后仰，眼前一黑，一口鲜血喷出。周芷若偷袭成功，左手跟着前探，五指便抓向他胸口，张无忌身受重伤，心神未乱，眼见这一抓到来，立时便是开膛破胸之祸，勉强向后移了数寸。嗤的一响，周芷若五指已抓破了他胸口衣衫，露出前胸肌肤

真实：周芷若右手五指跟着便要进袭，其时俞莲舟被她一腿踢倒，正中穴道，动弹不得，殷梨亭扑上要救援，也已不及，眼见张无忌难逃此劫。周芷若一瞥之下，忽然见到他胸口露出一个伤疤，正是昔日光明顶上自己用倚天剑刺伤的，五指距他胸膛不到半尺，心中柔情忽动，眼眶儿一红，竟然抓不下去。她稍一迟疑，韦一笑、殷梨亭、杨逍、范遥四人已同时扑到。韦一笑飞身挡在张无忌身前，杨范二人分袭周芷若左右，殷梨亭已抱着张无忌逃开

生成：上喝，想不到这句话竟是一人坏人，她想告诉她把她住手盘之成，我这般傲纵这件事，她若来问的时候，第一也已不会银票，你再看那一家伙骑马勿男，不管你这家伙使如猪狂的躯子，但空手？〞〝等到我那六个儿是干么？〞劳

德诺说：〝做父生，不得很多别别界的。〞那仇喝道：〝请请回来！」船弟子的

时候看得小乌的变故，在混过，这认得这么容易，根本不会想过，他老人英棺心趁他一条巷中还要撞人未见了，只要有甚麽兵刃、硼仙之罪，败以高升的高个人家来，因此紧急，是在天下皆派的少年的时，探窝。“我想决不成你看，也不会少了吧？你怎么好意，他要杀你。”红莲花道：“已经训示做！”语声中面首已不敢佩服，似乎能将一根联手，不是伊风四大镖局。

该段语句得分 0.543，属于平均水平，可以看到有点不知所云，只有偶尔的几句话是人话，但与上下文之间也没有明显联系。但是还是存在一定的亮点，就是每个人说的话基本都会被引号包含，而且可以明显发现这些话确实在一定程度上是以第一人称的语气说的。这说明我们的模型至少学到了一部分语言表达能力。

附录中还展示了评分最高和最低的语句，以供参考。

Conclusion

本次作业使用 LSTM 进行了文本生成，虽然最终效果较差，但还是能够一定程度的完成任务。个人认为如果能够按照词语建立词典，然后进行生成，应该可以取得更好的结果，但这也面临着词典无法包含所有词的风险。另一方面，还是需要更大的数据集，更大的模型才能更好完成任务。

Appendix

生成得分最高的语句，其得分为 0.842。

原文：洪七公右手持杯，左手拿着一只火腿脚爪慢慢啃着，说道：“常言道：物以类聚，人以群分。爱钱的财主是一帮，抢人钱财的绿林盗贼是一帮，我们乞讨残羹冷饭的叫化子也是一帮……”黄蓉拍手叫道：“我知道啦，我知道啦。那梁老怪叫你作‘洪帮主’，原来你是乞儿帮的帮主。”洪七公道：“正是。我们要饭的受人欺，被狗咬，不结成一伙，还有活命的份儿么？北边的百姓眼下暂且归金国管，南边的百姓归大宋皇帝管，可是天下的叫化儿啊……”黄蓉抢着道：“不论南北，都归你老人家管。”洪七公笑着点点头，说道：“正是。这根竹棒和这个葫芦，自唐末传到今日，已有好几百年，世世代代由丐帮的帮主执掌，就好像皇帝小子的玉玺、做官的金印一般。”黄蓉伸了伸舌头，道：“亏得你没给

我。〞洪七公笑问：〝怎么？〞黄蓉道：〝要是天下的小叫化都找着我，要我管他们的事，那可有多糟糕？〞洪七公叹道：〝你的话一点儿也不错。我生性疏懒，这丐帮帮主当起来着实麻烦，可是又找不到托付之人，只好就这么将就着对付了。

真实：黄蓉道：〝因此那梁老怪才怕得你这么厉害，要是天下的叫化子都跟他为难，可真不好受。每个叫化子在身上捉一个虱子放在他头颈里，痒也痒死了他。〞洪七公和郭靖哈哈大笑。笑了一阵，洪七公道：〝他怕我，倒不是为了这个。〞黄蓉忙问：〝那为了甚么？〞洪七公道：〝约莫二十年前，他正在干一件坏事，给我撞见啦。〞黄蓉问道：〝甚么坏事？〞洪七公踌躇道：〝这老怪信了甚么采阴补阳的邪说，找了许多处女来，破了他们的身子，说可以长生不老。〞黄蓉问道：〝怎么破了处女身子？〞黄蓉之母在生产她时因难产而死，是以她自小由父亲养大。黄药师因陈玄风、梅超风叛师私逃，一怒而将其余徒弟挑断筋脉，驱逐出岛。桃花岛上就只剩下几名哑仆。黄蓉从来没听年长女子说过男女之事，她与郭靖情意相投，但觉和他在一起时心中说不出的喜悦甜美，只要和他分开片刻，就感寂寞难受。她只知男女结为夫妻就永不分离，是以心中早把郭靖看作丈夫，但夫妻间的闺房之事，却是全然不知。她这么一问，洪七公一时倒是难以回答。黄蓉又问：〝破了处女的身子，是杀了她们吗？〞洪七公道：〝不是。一个女子受了这般欺侮，有时比给他杀了还要痛苦，有人说'失节事大，饿死事小'，就是这个意思了。〞黄蓉茫然不解，问道：〝是用刀子割去耳朵鼻子么？〞洪七公笑骂：〝呸！也不是。傻丫头，你回家问妈妈去。〞黄蓉道：〝我妈妈早死啦。〞洪七公〝啊〞了一声，道：〝你将来和这傻小子洞房花烛夜时，总会懂得了。〞黄蓉红了脸，撅起小嘴道：〝你不说算啦。〞这时才明白这是羞耻之事，又问：〝你撞见梁老怪正在干这坏事，后来怎样？〞洪七公见她不追问那件事，如释重负，呼了一口气道：〝那我自然要管哪。这家伙给我拿住了，狠狠打了一顿，拔下了他满头白发，逼着他把那些姑娘们送还家去，还要他立下重誓，以后不得再有这等恶行，要是再被我撞见，叫他求生不能，求死不得。听说这些年来他倒也没敢再犯，是以今日饶了他性命。他奶奶的，他的

头发长起了没有？〞黄蓉格的一声笑，说道：〝又长起啦！满头头发硬生生给你拔个干净，可真够他痛的了。〞三人吃过了饭。黄蓉道：〝七公，现下你就算把竹棒给我，我也不敢要啦，不过我们总不能一辈子跟你在一起。要是下次再碰见那姓梁的。他说：丶好，小丫头，前次你仗着洪帮主的势，用竹棒打我，今日我可要报仇啦。我拔光了你的头发！′那我们怎么办？先前靖哥哥跟这老怪动手，来来去去就只这么一招丶亢龙有悔′，威力无穷，果然不错，可不是太嫌寒蠢了些么？那老怪心里定是在说：丶洪帮主自己武功深不可测，教起徒儿来却是平平无奇。′

生成：〞阿朱道：〝我弄我亲眼睛？〞陆小凤道：〝赵快先生像一条，七十一拳记不可。〞楚留香叹了口气，喃喃道：〝其实我没有将他二字并不是他的本领，你也同样快了很奇怪，这个人立刻而不进车方？〞不懂他们信来对付他来到这三招，一步，我自己开始今天要让你妈妈呢？〞老人颔首阻住她了胆，他相信武三爷、华玉马的旧情马木洞门而去，迎面上官刃死下，而终不敢接到了这「黄河后的神气，俱经全有情网。田思思又吃惊道：〝她的确还可知谁呀！我又是个是丶等着谁要睡冷出的呻吟！一点点也被紧冒空。胡斐忙道：〝在下狱里？〞〝因为你若除掌后胜夷〞三字书生，却愉快越可疑，吓得第一两个会根本没有答。她，心想：〝胡说八〞突然一向他手下有一名′和悲护于他足，号令。〞不等他斟着一剑，〝为什么要爹爹露了一点手？〞陈家洛道：〝不明教主剑法。〞俞六家准备先行迹，这活龙枪上和地狱中夤夜遇见上官刃的父亲精为客友，过一个小孩，芮玮……〞陆有人笑嘻嘻的光淡，颔首道：〝方知不回盗的六年高强，岂知我却也是个大大人复仇，得过是的光花。至于深遂更难，那么杀？〞郭靖吃吃得太多反手道：〝原来这年大师和尚又生死不定而已。〞一风呼呼芝身上，更没有退进又有曲箭，寻思疑的惨呼，仍

生成得分最低的语句，其得分为 0.042。

原文：石清又是微微一笑，说道：“这吴道通跟我们素不相识，说不上得罪了愚夫妇什么。我们追寻此人，说来倒教周世兄见笑，是为了此人所携带的一件物事。

真实：周牧脸上肌肉牵动了几下，随即镇定，笑道：“贤夫妇消息也真灵通，这个讯息嘛，我们金刀寨也听到了。不瞒石庄主说，在下这番带了这些兄弟们出来，也就是为了这件物事。唉，不知是那一个狗杂种造的谣，却累得双笔吴道通枉送了性命。我们二百多人空走一趟，那也罢了，只怕安大哥还要怪在下办事不力呢。江湖上向来谣言满天飞，倘若以为那件物事真是金刀寨得了，都向我们打起主意来，这可不冤么？张兄弟，咱们怎么打死那姓吴的，怎样搜查那间烧饼铺，你详详细细的禀告石庄主、石夫人两位。

生成：”俞佩玉的冲天道：“曲浩子简单

~~~~~~
~~~~~~
~~~~~~
~~~~~~
~~~~~~
~~~~~~
~~~~~~
~~~~~~
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~