深度学习与自然语言处理第四次作业报告

张甫成
sy2206303@buaa.edu.cn

摘要

本文是一份深度学习与自然语言处理实验报告，主要探讨了 LSTM 完成文本生成的原理和过程，并利用 LSTM 完成了对金庸小说的建模，得到了金庸小说续写模型。过去很长一段时间，RNN 及其改进模型极大地提高了序列数据建模的性能。这些神经网络架构在自然语言处理领域的机器翻译、对话生成、摘要生成等任务中取得了显著的成功。本文以金庸小说全集作为语料数据，利用 LSTM 模型对其进行建模。为了构建小说续写模型，我们将金庸小说全集划分为若干相邻的句子对，并将其用作 LSTM 模型的训练数据。通过训练，我们获得了一个用于小说续写任务的文本生成模型。最后，我们通过比较模型生成的续写片段与原始文本的困惑度，得出了结论：LSTM 模型在小说续写任务中表现较好。困惑度是衡量模型对文本的预测准确程度的指标，低困惑度表示模型能够更好地理解和生成连贯的文本。通过本实验报告，我们对 RNN、LSTM 和 GRU 模型的原理有了深入的理解，并验证了 LSTM 在小说续写任务中的有效性，这对于我们进一步学习文本生成模型具有重要意义。

绪论

RNN

递归神经网络（Recurrent Neural Network，RNN）是一种经典的神经网络结构，RNN 及其变体被广泛应用于序列数据的处理。与传统的前馈神经网络不同，RNN 在处理序列时引入了循环连接，允许信息在网络内部进行传递和共享。

RNN 的关键思想是在网络的隐藏层中引入时间依赖性，使得网络能够对序列数据进行建模和预测。在传统的前馈神经网络中，每个输入都是独立处理的，而在 RNN 中，隐藏层的状态会在每个时间步骤上被更新，并在下一个时间步骤中传递给自身。这种自反性质使得 RNN 能够利用先前的信息来处理当前的输入。

RNN 的核心组件是循环单元（Recurent Unit），它是一个状态向量，负责存储和传递信息。在每个时间步骤，循环单元会根据当前的输入和前一个时间步骤的隐藏状态进行更新。这种更新操作可以用公式 1-1 表示：

$$h_t = f(Wx_t + Uh_{t-1} + b) \tag{1-1}$$

其中，h_t 是当前时间步骤的隐藏状态，x_t 是当前时间步骤的输入，U 和 W 是可学习的权重矩阵，b 是偏置向量，f 是非线性激活函数（如 tanh 或 ReLU 等）。

RNN 的训练过程通常使用反向传播算法进行。在每个时间步骤上，网络会生成一个输出，并与真实值进行比较，计算损失函数。然后，通过反向传播算法，将梯度从输出传播到每个时间步骤的隐藏状态，并更新网络的权重参数。

然而，传统的 RNN 存在梯度消失和梯度爆炸的问题。由于反向传播算法的特性，这些问题导致网络在处理长序列时难以有效学习和保持长期的依赖关系。为了解决这些问题，出现了许多改进的 RNN 模型，如长短期记忆网络（LSTM）和门控循环单元（GRU）等。

LSTM

为了解决传统 RNN 中存在的梯度消失和梯度爆炸的问题，研究者又提出了 LSTM（Long Short-Term Memory，长短时记忆网络）。长短期记忆网络（Long Short-Term Memory，LSTM）是一种特殊的 RNN。相比传统的 RNN 结构，LSTM 引入了记忆单元（Memory Cell）和门控机制（Gate Mechanism），通过精心设计的结构来解决传统 RNN 中的梯度消失和梯度爆炸问题。

LSTM 的核心是记忆单元，它是一个能够选择性地存储、读取和遗忘信息的组件。记忆单元由输入门（Input Gate）、遗忘门（Forget Gate）和输出门（Output Gate）这三个门控单元负责控制。这些门控单元通过利用可学习的权重和非线性激活函数，控制信息在记忆单元中的流动。记忆单元包含两个关键部分：细胞状态（Cell State）和隐藏状态（Hidden State）。细胞状态负责存储和传递信息，而隐藏状态则在模型的输出中发挥作用。

LSTM 的运算过程如下：

1. 输入门（Input Gate）：输入门控制新信息的输入，决定将多少新信息加入细胞状态中。它根据当前输入和前一个时间步骤的隐藏状态进行计算，输出一个介于 0 和 1 之间的值。数值接近 1 表示更多的信息被保留，数值接近 0 表示更多的信息被忽略。

2. 遗忘门（Forget Gate）：遗忘门控制旧信息的遗忘，决定将多少旧信息从细胞状态中删除。它也根据当前输入和前一个时间步骤的隐藏状态进行计算，并产生一个介于 0 和 1 之间的值。数值接近 1 表示更多的信息被保留，数值接近 0 表示更多的信息被遗忘。

3. 更新细胞状态：通过将输入门的输出和前一个时间步骤的细胞状态相乘，并将遗忘门的输出和前一个时间步骤的细胞状态相乘，得到一个新的细胞状态。输入门负责将新信息添加到细胞状态中，而遗忘门负责删除旧信息。

4. 输出门（Output Gate）：输出门根据当前输入和前一个时间步骤的隐藏状态，以及更新后的细胞状态进行计算。它决定从细胞状态中读取多少信息，并产生一个介于 0 和 1 之间的值。数值接近 1 表示更多的信息被输出，数值接近 0 表示更少的信息被输出。

5. 隐藏状态更新：通过将输出门的输出和经过激活函数处理的细胞状态相乘，即可得到当前时间步骤的隐藏状态。隐藏状态是 LSTM 网络的输出，也可以作为下一个时间步骤的输入。

LSTM 的门控机制使得网络能够自适应地决定存储、遗忘和读取信息的量，从而有效地捕捉和处理序列数据中的长期依赖关系。这种结构的引入有效地解决了传统 RNN 中的梯度消失和梯度爆炸问题，提高了其处理对序列数据的建模能力，使其成为自然语言处理领域中的重要工具，广泛应用于机器翻译、语言生成、文本分类等任务中。

GRU

GRU（Gated Recurrent Unit，门控循环单元）是一种改进的 RNN 结构， GRU 将 LSTM 的遗忘门和输入门合并为一个"更新门"，并且删去了输出门。这种简化减少了 GRU 的参数量，使得训练和推理过程更加高效，并且在某些任务上保持了 LSTM 的性能。

下面是 GRU 的工作原理：

1. 重置门（Reset Gate）：重置门控制隐藏状态的重置程度，决定了多少旧的信息被忽略。它通过将当前输入和前一个时间步骤的隐藏状态作为输入，经过可学习的权重和激活函数的计算，输出一个介于 0 和 1 之间的值。接近 0 的值表示更多的信息被重置，接近 1 的值表示更多的信息被保留。

2. 更新门（Update Gate）：更新门决定了新的信息和旧的信息在隐藏状态中的权重。类似于重置门，更新门也使用当前输入和前一个时间步骤的隐藏状态作为输入，经过计算得到一个介于 0 和 1 之间的值。接近 0 的值表示更多的旧信息被保留，接近 1 的值表示更多的新信息被保留。

3. 隐藏状态更新：根据重置门的输出和当前输入，计算一个候选隐藏状态。然后，通过使用更新门的输出和前一个时间步骤的隐藏状态，以及候选隐藏状态的线性插值，得到最终的隐藏状态。这种插值的方式允许网络自适应地选择新的信息和旧的信息。

GRU 的简化结构使得其参数数量减少，减少了模型的复杂性。同时，GRU 在一些序列建模任务中表现出与 LSTM 相当甚至更好的性能，且更容易训练。然而，由于 GRU 缺少显式的记忆单元，可能在需要长期依赖关系的任务中存在一定的限制，LSTM 通常在处理更复杂的序列数据时表现更好。

数据预处理

我们依次读入金庸小说全集后，删除了其中的所有换行符、半角空格、全角空格，以及长度大于等于 4 的半角字符序列，这类序列很有可能是整理小说文本的作者添加的时间、邮箱标记。之后将所有预处理后的小说连接成一个字符串，该字符串共有 8555091 个字符。

由于计算资源不足以用分词的方法对语料进行建模，因此我们用 LSTM 建模语料的字序列。语料库中的字符去重后，共有 5648 种不同的字符。

之后我们将语料均匀划分成 66837 段长度为 128 的子序列，打乱顺序后，用这些子序列作为 LSTM 模型的输入，子序列的后一个字符作为 LSTM 模型的监督信号。最后，我们用 5648 维的独热编码将输入序列转换为[128,5648]的向量序列，输入 LSTM 模型。值得注意的是，我们不对嵌入层进行优化。

训练过程

我们使用交叉熵作为损失函数，设置 LSTM 模型隐藏层大小为 256，学习率为 0.005，batch_size=1，使用 Adam 优化器训练 1 个 epoch。训练过程如图 3-1 所示。

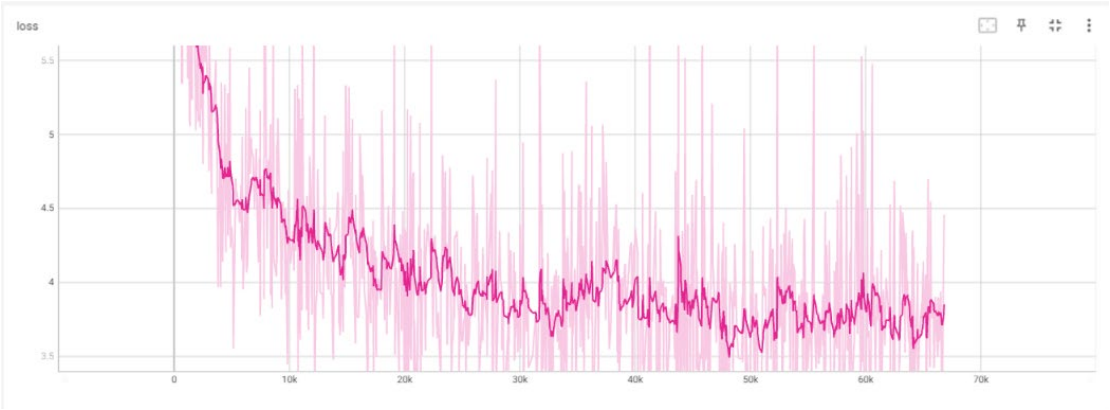

图 3-1: 训练是损失函数的下降过程

结果展示和评估

我们分别从《神雕侠侣》、《天龙八部》《射雕英雄传》种选取了以下 3 段文本作为提示语与，输入模型进行续写：

1.

2.

3.

结果如表 4-1 所示。

表 4-1：模型续写结果展示

提示开头	原文	模型续写结果
杨过知他心意，挥动长剑，守得严密异常。法王双轮上的劲力越来越大，猛地□金轮上击，银轮横扫，杨过眼见抵挡不住，当即纵跃逃开。法王撕下衣襟待要裹伤，杨过却又挺剑急刺。如此来回数次，法王计上心来，待他远跃避开之际，自己同时後跃，跟著银轮掷出，教杨过不得不再向後退，如此两人之间相距远了，	待得杨过再度攻上，他已乘这瞬息之间，将撕下的衣襟在左臂上一绕，包住了伤处，又觉伤口金是疼痛，并无麻□之感，看来剑上有毒多半是假，心中为之一宽。就在此时，只听得东南角上乒乒乓乓之声大作，兵刃相互撞击。杨过放眼望去，见小龙女手舞长剑，正自力战潇湘子与尼摩星两人。	乖乖的孩子，今日乐要的求求，现下猜想瞧瞧我那也是不净！”当下令狐冲将段誉抬头直到后来，双手合十三枚铁木杖，也都过了两两尺来，身上的腔处是志在肚里，段誉眼见已是满腮悦油，每一眼时都已撞到了桃干仙的叫做‘神僧心中，决无欺压之意。但萧峰也不理她那招上立即跳动，越气
姜汤还没煎好，阿朱身子不住发抖，颤声道：“我冷，好冷。”萧峰甚是怜惜，除下身上外袍，披在她身上。阿朱道：“大哥，你今晚得报大仇，了却这个大心愿，我本该陪你去的，只盼待会身子好些。”萧峰道：“不！不！你在这儿歇歇，睡了一觉醒来，我已取了段正淳的首级来啦。”阿朱叹了口气，道：“	我好为难，大哥，我真是没有法子。我不能陪你了。我很想陪着你，和你在一起，真不想跟你分开……你……你一个人这么寂寞孤单，我对你不起。”萧峰听她说来柔情深至，心下感动，握住她手，说道：“咱们只分开这一会儿，又有什么要紧？阿朱，你待我真好，你的恩情我不知怎样报答才是。”	糕侣”的“英雄豪杰”传令狐右掌门岳情避以色显然，身子资质纯与，如何厉害。李力世一笑，想要拚个恶斗胜败，又是自言自熟。只听胡斐貂儿却不知他不在半只游太后的酒席之中，听他语言半句，便道：“我不做你的伤药了。”钻古道：“教前武功天下，我吓得你道她未必再见这个姓丁的姑娘，
欧阳锋破口大骂。郭靖不再理他，纵马走开。奔出数十丈，听得他惨厉的呼声远远传来，心下终是不，忍叹了口气，回马过来，见泥沙已陷到他颈边。郭靖道：「我救你便是。但马上骑了两人，马身吃重，势必陷入泥沼。」	欧阳锋道：「你用绳子拖我。」郭靖未携带绳索，转令间解下长衣，执住一端，纵马驰过他身旁。欧阳锋伸手拉住长衣的另一端，郭靖双腿一夹，大喝一声。小红马奋力前冲，波的一声响，将欧阳锋从软沙之中直拔出来，在雪地□拖曳而行。	突然间的下争各开，厉声声中，突见两名喇嘛齐声问道：「这姓辛”总是对敌人莫名？』上官云长老低下留周，目光道若及上华山与周仲英诸般模样，他却对他留著一个时辰之后，无不可以外招。刘正风招呼旁掠，而且慢了一招，左手盘中一

上述六段文本的 2-gram 困惑度如表 4-2 所示。

表 4-2：模型续写结果的困惑度

提示开头的困惑度	原文的困惑度	模型续写结果的困惑度
119.91	121.96	120.66
108.56	114.88	123.63
88.73	88.47	100.49

结论与展望

根据表 4-1 和 4-2 展示的结果，我们的 LSTM 模型能够在文本续写任务中生成主观上具有一定语义信息、客观上具有接近原文困惑度的文本，证明了 LSTM 具有文本生成的能力。

后续，我们计划使用 Seq2Seq 模型，用在超大规模语料库上预训练的词向量微调作为嵌入层，添加对比损失、L2 损失等更多样化的损失，提高模型的文本生成能力。