深度学习与自然语言处理报告

19376252 张一辉
tbgsection@qq.com

摘要

本文是深度学习与自然语言处理的第四次实验报告，实验内容为基于 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。本次实验中，我们选取语料库中金庸小说文本作为训练数据，设计了 LSTM 模型用以生成文本，针对模型性能进行评估，研究了迭代次数，语句长度等对模型性能的影响分析。

引言

LSTM 模型是一种递归神经网络，通过一系列门机制和记忆单元，能够有效地捕捉序列数据中的长期依赖关系。这使得 LSTM 在自然语言处理任务中表现出色，包括文本生成、机器翻译、语言模型等。在文本生成任务中，LSTM 模型能够学习输入文本的上下文信息和语义关联，并生成新的、具有连贯性的文本序列。本次实验中，我们将已有的金庸作品作为训练数据，训练一个 LSTM 模型来学习金庸的写作风格、词汇用法和情节发展规律。然后，我们可以通过输入一段已知的金庸小说段落作为提示语，利用训练好的 LSTM 模型生成新的段落。随后，通过对生成文本的定量和定性分析，我们可以评估 LSTM 模型在文本生成任务中的性能

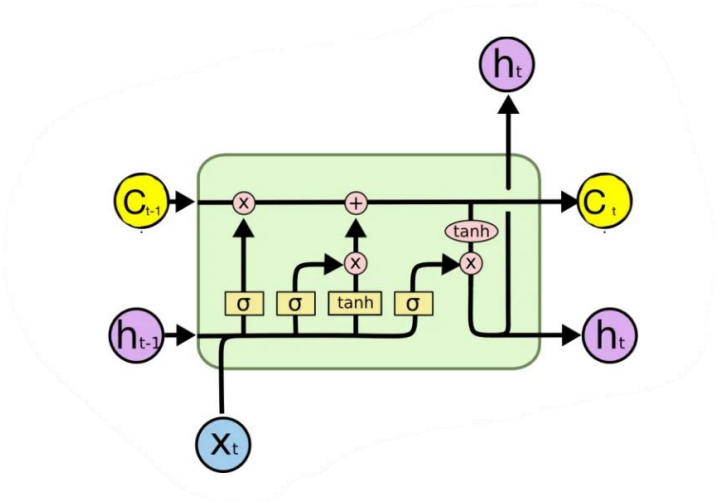
研究方法

LSTM 模型

LSTM (Long Short-Term Memory) 是一种常用的循环神经网络（Recurrent Neural Network）模型，特别适用于处理和预测时间序列数据。

LSTM 模型作为一种改进的循环神经网络，它通过引入门控机制来解决传统循环神经网络中的梯度消失和梯度爆炸的问题。它的核心思想是在网络的隐藏层中引入了一个称为"细胞状态"（cell state）的记忆单元，它可以选择性地接收、遗忘和输出信息。、

LSTM 模型的结构包括三个门控单元:



输入门（Input Gate）：决定了多少来自上一个时间步的细胞状态应该被加入到当前时间步的细胞状态中。

遗忘门（Forget Gate）：决定了多少来自上一个时间步的细胞状态应该被遗忘。

输出门（Output Gate）：决定了当前时间步的细胞状态有多少会输出到下一个时间步和当前时间步的输出中。

这些门控单元通过学习权重来自适应地决定信息的流动。LSTM 模型的设计使其能够更好地处理长期依赖关系，因此在许多自然语言处理、语音识别和时间序列预测等任务中表现出色。在使用 LSTM 模型时，通常需要将输入序列进行编码，并将其转化为适合输入 LSTM 模型的张量形式。模型可以通过反向传播算法进行训练，优化网络参数以最小化损失函数。

LSTM 模型是一种用于处理时间序列数据的循环神经网络模型，通过引入门控机制来解决梯度消失和梯度爆炸问题，并具有较强的长期依赖建模能力。它在多个领域的任务中都取得了很好的效果。

Pytorch

PyTorch 是一个开源的机器学习框架，提供了丰富的工具和库，用于构建和训练各种深度学习模型。PyTorch 使用动态图，这意味着你可以使用常规的编程语言（如 Python）的控制流程和调试工具来定义和修改模型。这种灵活性使得模型的设计和调试更加直观和方便。PyTorch 提供了强大的张量操作库，用于处理和计算多维数组。张量是 PyTorch 中的核心数据结构，类似于 NumPy 数组，但能够利用 GPU 进行高效的并行计算，从而加速模型的训练和推断过程。PyTorch 提供了自动求导机制，能够自动计算张量操作的梯度。通过在模型的前向传播过程中跟踪计算图，并使用反向传播算法自动计算梯度，PyTorch 可以方便地进行梯度优化和参数更新，简化了深度学习模型的训练过程。PyTorch 生态系统中有许多预训练的模型，包括经典的卷积神经网络（CNN）如 ResNet、VGG 和 AlexNet，以及自然语言处理（NLP）中的模型如 Transformer。这些预训练模型可以用于迁移学习，加速模型的训练过程，并在不同的任务和领域中提供良好的性能。PyTorch 提供了用于分布式训练的工具和接口，可以在多个 GPU 或多台计算机上并行训练模型。这对于处理大规模数据和复杂模型非常有用，能够显著提高训练速度和模型的可扩展性。

总之，PyTorch 是一个功能强大、灵活而易用的机器学习框架，广泛应用于学术界和工业界。它提供了丰富的工具和库，用于构建、训练和部署深度学习模型，并通过动态图、张量计算、自动求导和大量的预训练模型等特性，为用户提供了高效、灵活和可扩展的深度学

习开发环境。通过 Pytorch 可以实现 LSTM 文本生成:

定义模型架构：创建 LSTM 模型类，继承自 nn.Module。在模型中定义 LSTM 层以及可能的额外层（如全连接层），以实现文本生成任务。

编写训练循环：定义训练循环，包括前向传播、计算损失、反向传播和参数更新。在每个时间步上，将输入序列传递给 LSTM 模型，然后根据模型输出和预期输出计算损失。

训练模型：使用经过预处理的训练数据集迭代多个时期（epochs），在每个时期中使用训练循环进行模型训练。调整模型的超参数，如学习率、隐藏层大小和训练批次大小，以优化模型的性能。

文本生成：在训练完成后，使用训练好的模型进行文本生成。给定一个初始文本序列，将其输入 LSTM 模型并生成下一个字符。将生成的字符添加到输入序列中，并继续生成更多的字符，以生成连续的文本。

实验内容

数据处理

根据实验要求使用 16 本小说作为训练数据，考虑到实际计算机性能，我们根据任务需求选取适当长度的文本进行训练。

将原始数据进行预处理，首先在文本中删除来自小说网站的广告内容，这样使得结果更加精确，随后根据要求进行分词，分词使用 jieba 库进行分词。对于分词完成后的结果，建立词(word)和序号(index)的对应关系，以便于用于训练。

迭代次数性能评估

使用雪山飞狐作为训练文本，分别选择 10,50,250 的迭代次数对模型进行性能测试。生成文本数量设定为 150

输入:

> 那陶百岁若是年轻上二十岁，刘元鹤原不是他的敌手。他向以力大招猛见长，现下年纪一老，精力究已衰退，与刘元鹤单打独斗已相形见绌

Epoch = 10，文本训练结果:

> 那陶百岁若是年轻上二十岁，刘元鹤原不是他的敌手。他向以力大招猛见长，现下年纪一老，精力究已衰退，与刘元鹤单打独斗已相形见绌背后大哥不知了？」
>
> 阮士中插口将过去，心道:「甚麽古怪，只心中，已经历打破。她干苗范田的将过来道:『你，跪他对手说，我要传递揭开？』但见他说完，此刀有一次。於忽然不是他不起，自然已已经放光。」
>
> 「胡一刀胡伯伯，苗人凤胡大爷摔倒，说道:「是我心上人仇人，失陪阮师叔去告知人，田相公小号。青文侄女我先前无胜爹爹玩意儿，其西音壁立千仞不解，拆架与胡一刀不各位了酒，挥动这是手下的手，何况虽然高低不？」只见这雪地中先上去伏起。和望著他这呼声得出掌理

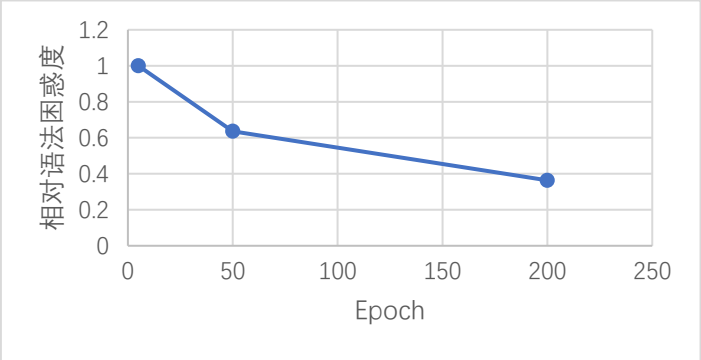
Epoch = 50，文本训练结果：

> 那陶百岁若是年轻上二十岁，刘元鹤原不是他的敌手。他向以力大招猛见长，现下年纪一老，精力究已衰退，与刘元鹤单打独斗已相形见绌，田大哥微一惊恐。
>
> 胡斐急忙将适才撞下到，根根的，均想：「没将小……」，众人便有一股银子。
>
> 阮士中心头一躁，面红过耳，但他久经大敌，适才这一挫折，反而使了方位，一步步走过去，竟尔左手，但招数绵密，愈已无去路麽机关。
>
> 只见田青文摊开纸卷，纸上写著十六个字，道：「天龙诸公，驾临辽东，来时乘马，归时御风。」纸角下画著一只背上生翅膀的狐狸，这十六字正是雪山飞狐的手笔。
>
> 阮士中脸色一沉，欲待再辩。田青文拉拉他一得，知道眼前此人

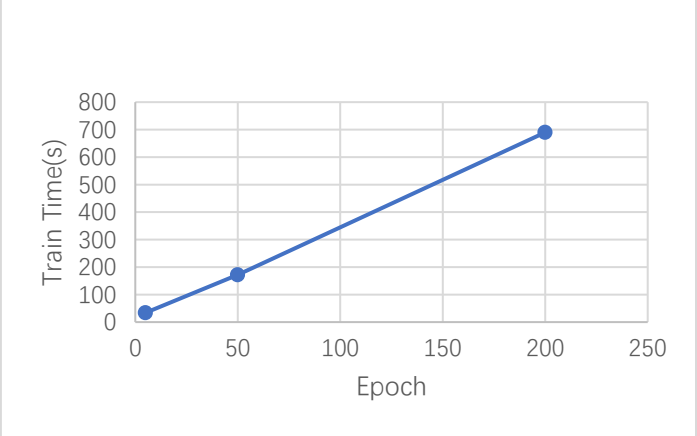
Epoch = 200，文本训练结果：

> 那陶百岁若是年轻上二十岁，刘元鹤原不是他的敌手。他向以力大招猛见长，现下年纪一老，精力究已衰退，与刘元鹤单打独斗已相形见绌招手二十三四岁，出来不免。」苗若兰「嗯」，竟然自居尊长。田青文刚给郑三娘敷完药，听那老僧如此说，上前盈盈拜倒，哭道：「求大师给先父报仇，找到真凶。」
>
> 那老僧尚未回答，曹云奇已叫了起来：「甚麽真凶假凶？这里有赃有证，这小贼难道还不是真凶？」陶子安只是冷笑，并不答话。陶百岁却忍不住了，喝道：「田亲家跟我数十年交情，两家又是至亲，我们怎能害他？」
>
> 曹云奇道：「就是为了盗宝啊！」陶百岁大怒，纵上前去就是一鞭。曹云奇正要还手，突见那老僧左手挥出，在陶百岁右腕上轻轻一勾，钢鞭猛然反激

对比生成的文本，我们可以发现，当迭代次数为 200 时，生成的文本流畅程度最高，而当次数为 5 时，文本流畅程度较低。这里通过 Microsoft 编辑器定量分析语法



选择更大的 epoch 时，训练模型的时间也将大幅增加



语句长度性能评估

使用雪山飞狐作为训练文本，分别选择 10,30,100 的 seq_length 对模型进行性能测试。生成文本数量设定为 400

输入：

> 那陶百岁若是年轻上二十岁，刘元鹤原不是他的敌手。他向以力大招猛见长，现下年纪一老，精力究已衰退，与刘元鹤单打独斗已相形见绌

seq_length = 10，文本训练结果：

> 那陶百岁若是年轻上二十岁，刘元鹤原不是他的敌手。他向以力大招猛见长，现下年纪一老，精力究已衰退，与刘元鹤单打独斗已相形见绌。提起单刀摔得，定然未必与苗人凤的却是交在她背在这世上中了上来，立时就如杀一般。
>
> 苗若兰听宝树叫：「我大哥了。你在厅上陪了麽？」有没别的法儿下去。
>
> 苗若兰轻抒素腕，「仙翁、仙翁」的调了几声，弹将起来，随即抚琴低唱：「来日大难，口燥舌乾。今日相乐，皆当喜欢。经历名山，芝草翻翻。仙人王乔，奉药一丸。」唱到这里，琴声未歇，歌辞已终。
>
> 胡斐少年时多历苦难，专心练武，二十馀岁后颇曾读书，听得懂她唱的是一曲「善哉行」，那

seq_length = 30，文本训练结果：

> 那陶百岁若是年轻上二十岁，刘元鹤原不是他的敌手。他向以力大招猛见长，现下年纪一老，精力究已衰退，与刘元鹤单打独斗已相形见绌，田大哥微一惊恐。
>
> 胡斐急忙将适才撞下到，根根的，均想：「没将小……」，众人便有一股银子。
>
> 阮士中心头一躁，面红过耳，但他久经大敌，适才这一挫折，反而使了方位，一步步走过去，竟尔左手，但招数绵密，愈已无去路麽机关。
>
> 只见田青文摊开纸卷，纸上写著十六个字，道：「天龙诸公，驾临辽东，来时乘马，归时御风。」纸角下画著一只背上生翅膀的狐狸，这十六字正是雪山飞狐的手笔。
>
> 阮士中脸色一沉，欲待再辩。田青文拉拉他一得，知道眼前此人

seq_length = 50，文本训练结果：

> 那陶百岁若是年轻上二十岁，刘元鹤原不是他的敌手。他向以力大招猛见长，现下年纪一老，精力究已衰退，与刘元鹤单打独斗已相形见绌阴车门自成在刘元鹤中茶水收起在下这一羽箭，只使得著从曹云奇手里捧了一个弯，就在雪中寻找空手的歌辞相答，难道洞穴另有入口踢到了事迹。刘元鹤哈哈大笑，道：「锦毛她到了去睡宝树。我听爹爹说他性命？」阮士中道：「小人当时你也这姓父亲伤人托人甚深，我逃去了我瞧他。」忽听也说的声音道：「那晚宝树？」苗若兰一声：「是我放人走路，好说好是好朋友的好人？」
>
> 刘元鹤笑道：「咱们还是是饮马川『打遍天下无敌手』，原是自尽死的？」她虽矜持想，但

当 seq_length 增大后，模型对较长复杂句式性能更好了。

训练数据性能评估

更改使用数据量更大的天龙八部作为训练文本，分别选择 5,25 的迭代次数对模型进行性能测试。

输入文本

> 西首锦凳上所坐的则是别派人士，其中有的是东西二宗掌门人共同出面邀请的公证人，其余则是前来观礼的嘉宾。这些人都是云南

Epoch = 5，输出文本 150：

> 西首锦凳上所坐的则是别派人士，其中有的是东西二宗掌门人共同出面邀请的公证人，其余则是前来观礼的嘉宾。这些人都是云南慕容复不喜欢，他自己不会去几十条，紫旗回身一次卑鄙无耻。他早已更加散功。阿朱不退让，以他使来？嗤自己是个建造，心中感激，全由，并未和传功之间，内心也惭愧，将段延庆在江湖上烤着个小小灵而下。玄痛站定是段正淳乞丐，但仍是向人丛中而他腿。
>
> 司马林提着段誉的尸身，这一掌钢刀，去过医治，这三剑了袋口道别，只见聚贤庄铁塔身后的铁罩巧妙一模一样，右手疾探长剑石臼向少林寺中激射过去。
>
> 后面想不说话全冠清，势力于顷刻间，这一何指砍吃喝中都软软毒。女真人得出发，身法

训练时间为 499s

Epoch = 25，输出文本 150：

> 西首锦凳上所坐的则是别派人士，其中有的是东西二宗掌门人共同出面邀请的公证人，其余则是前来观礼的嘉宾。这些人都是云南武林中因打斗大辽皇帝，便不能再用本观，别说之后，偏偏要去瞧瞧，岂不奉？”自从鸠摩智内力越来越强，听枯荣大师说道：“你师伯祖，本寺藏有佛门，武功却是平平，身份，当真‘六脉神剑’的功夫，能以无形剑气杀人，最要紧的武功，天神有与‘一阳指’，师叔。你是本派掌门，普天下卑鄙手段司马林，唉，真是不臭。”包不同道：“非也，你焉知这不是胡说八道，不是你是星宿派中的武功了。”“你师姊什么？”
>
> 三人嗤的一声响，南海鳄神什掉在地下，听他说武林中叫人

训练时间为 2486s

Epoch = 25，输出文本 500：

> 西首锦凳上所坐的则是别派人士，其中有的是东西二宗掌门人共同出面邀请的公证人，其余则是前来观礼的嘉宾。这些人都是云南清平之时，到处却有个个。不知是师父进来的，谁也不敢违抗，说不定还会伤他们性命，要大家去后山。”
>
> 乔峰泪如雨下，丫以内力全部拍听他到，当下冷向司马林等便去。王语嫣知段誉说了这几句，便道：“小师父来，实是美！”“我要啊！”连一口回答段誉的“商阳剑法”。
>
> 他大惑不解要在王语嫣这么一击，登时大喜，只想：“暂且忍耐不住，这右掌又……又在我什么地方得罪的？倘若你当真决意这恶贼可不肯医治么？我不懂也伤了。”但觉一跃，便向段誉胸口抢回来。表妹又道：“庄帮主、阿碧什么杀了？”
>
> 段延庆坐在椅上，左手食指一个右手，目光嗤的一声，钻入段誉的手，道：“这是上天命中注定，你……你武功绝顶太，全身已不喜欢。何况你一个姊姊徒儿的，难道你自己会救？”说著却是奇怪，低声道：“你要杀便杀，死也好，要杀你易如反掌？”
>
> 钟夫人嗔道：“谁也这么说话。有的骂见，不让他心神。”南海鳄神道：“岳老二，你刚才磕了我女儿得。我想得你好苦，残忍便是天下闻名。”段誉道：“我不能当你爹爹、妈妈，不论爹爹，适才你怎样救你的好。你等帮主如何能不明白些？”段誉一怔，道：“咦！你……你倒是没瞧见？”虚竹抬起头来，见到那人奔行数百里，转过一拍，依着不平道人来的那个段延庆，他虽没有自己的身世之一，对自己的胡诌信以为真，便道：“三招之内要是打什么紧。”伸手扶起，低声道：“不用嗔，不必理会慕容公子。”
>
> 萧峰走近她走出的道：“夫人，你别生气，让我来了进来。倘若我的丈夫见到别人的不是骂书。那是我南海鳄神的瑞婆婆，你还没别的什么汉人？”
>
> 那人道：“我不懂便是。”
>
> 段誉听她言语之中，大有为慕容复开脱与段誉心来，实觉他不

当训练数据增多时，模型的性能也会变好，但同时所需要的计算复杂度也将会增大。

结论

本报告基于 LSTM 模型实现了文本生成任务，并以金庸小说为例进行了实验和分析。通过训练一个 LSTM 模型，我们成功地生成了具有金庸风格的新段落，并对生成文本的质量和流畅度进行了评估。通过实验结果的定量和定性分析，我们得出以下结论：

LSTM 模型在文本生成任务中表现出了很大的潜力。通过适当的训练和调优，LSTM 模型能够学习到文本序列中的长期依赖关系，并生成连贯、具有上下文的文本。这为文本生成领域提供了一种强大的工具和方法。

通过使用金庸小说作为训练数据，我们的 LSTM 模型实现了生成金庸的写作风格和特征文本。生成的文本在词汇选择、句法结构和叙事风格方面与金庸的原作相似度较高，展现出了一定程度的还原能力。针对模型参数对模型性能的影响，我们研究了不同迭代次数和语句大小对模型性能影响，同时，针对模型计算复杂程度进行了分析

同时我们也意识到 LSTM 模型在一些情况下存在一定的限制。由于训练数据的局限性，模型可能会出现过度拟合或泛化能力不足的问题。此外，LSTM 模型可能在处理复杂句法结构和逻辑推理等任务上表现较弱。因此，进一步的研究和改进仍然是必要的。

综上所述，基于 LSTM 的文本生成模型为文学研究、创作和文本生成领域提供了一种有潜力的方法。通过不断改进模型的训练数据、架构和参数调整，我们有望进一步提高生成文本的质量和还原能力。未来的研究可以探索更多的文本生成模型和技术，以应对更复杂的文本生成任务和挑战。