Report of Deep Learning for Natural Langauge Processing

JiaXiang Yu

Zy2403124@buaa.edu.cn

Abstract

本研究利用给定语料库，用 LSTM 与 Transformer 两种不同的模型来实现文本生成的任务（给定开头后生成武侠小说的片段或者章节），并对比与讨论两种方法的优缺点。

Introduction

长短期记忆（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。简单来说，就是相比普通的 RNN，LSTM 能够在更长的序列中有更好的表现。

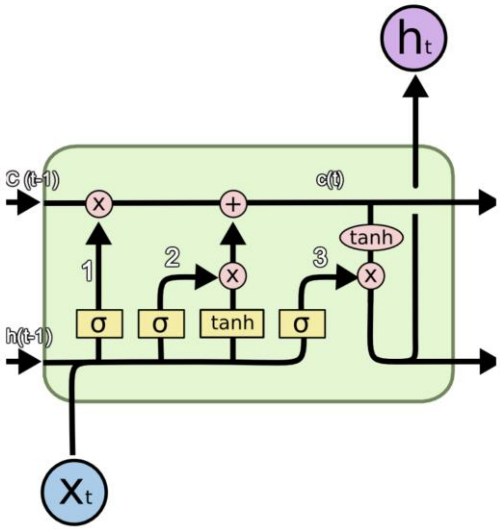

图 1 LSTM 示意图

Transformer 模型是由 Google 在 2017 年提出的，旨在解决传统的序列到序列模型在处理长距离依赖问题上的不足。传统的 RNN 和 LSTM 模型在处理长文本序列时，容易出现梯度消失或爆炸问题，导致模型性能下降。Transformer 模型通过引入自注意力机制和多头注意力机制，成功地解决了这一问题。

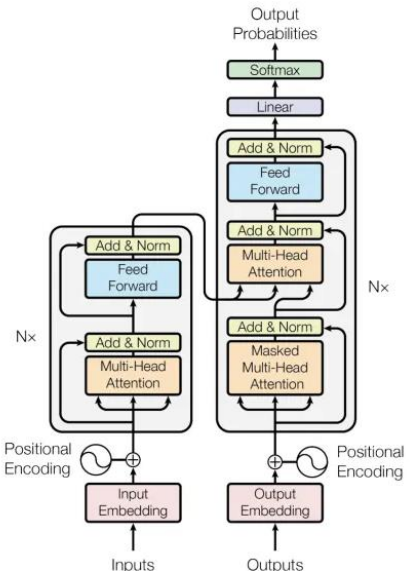

图 2 Transformer 示意图

Methodology

Step1: 基于LSTM的seq2seq模型建立

首先进行数据预处理，读取所有文本内容，去除无用的标点符号等杂项信息，并使用 jieba 库对文本进行分词。接着，定义一个自定义的数据集类，用于加载和处理分词后的小说句子，其中包含 collate_fn 函数以在批处理时对句子进行填充，确保长度一致。随后，统计词频并构建词汇表，加入特殊标记如 <pad>、<sos>、<eos> 和 <unk>。在模型结构方面，首先定义编码器（Encoder），其将输入序列嵌入到高维空间后，通过 LSTM 模块处理并输出隐藏状态和细胞状态。然后定义解码器（Decoder），它接收编码器传来的隐藏状态和细胞状态，并结合前一个时间步的输出，生成当前时间步的预测结果。接着定义整个 Seq2Seq 模型，将编码器和解码器组合起来，实现从输入序列到输出序列的转换，并通过 teacher forcing 机制决定是否使用真实目标输入还是模型预测结果作为下一个时间步的输入。在训练阶段，通过梯度累积来优化模型参数以适应小批量数据训练的需求。最后，模型可以根据给定的起始文本生成新的文本内容，完成文本生成任务。

Step2: Transformer模型的建立

首先，选择更适合中文语言建模的预训练模型，即 Hugging Face 模型库中的 “uer/gpt2-chinese-cluecorpussmall”，该模型在大规模中文语料 CLUECorpusSmall 上进行预训练，具备良好的中文语义理解与生成能力。随后，

下载所需的模型相关文件，包括模型结构配置文件 config.json、预训练权重文件 pytorch_model.bin 以及中文词汇表文件 vocab.txt，为模型的加载与使用做好准备。接着，通过 Hugging Face 的 Transformers 库加载模型和分词器，使其能够正确地对输入文本进行编码，并将模型输出解码为自然流畅的中文句子。在生成过程中，支持设置温度（temperature）、前缀（prompt）、最大长度（max_length）等参数，以控制生成文本的多样性与连贯性。通过调用模型的生成接口，可以根据给定的起始文本生成语义连贯、符合上下文逻辑的中文文本，实现高质量的文本生成任务。

Experimental Studies

经过实验评估，得到结果如下：

在 epoch=10，max_length=300 的条件下，实验结果如下：

给定的初始文本：青衣剑士回剑侧身，右腿微蹲，锦衫剑士看出破绽，挺剑向他左肩疾刺。不料青衣剑士这一蹲乃是诱招，长剑突然圈转，直取敌人咽喉，势道劲急无伦。锦衫剑士大骇之下，长剑脱手，向敌人心窝激射过去。这是无可奈何同归于尽的打法，敌人若是继续进击，心窝必定中剑。当此情形，对方自须收剑挡格，自己便可摆脱这无可挽救的绝境。不料青衣剑士竟不挡架闪避，手腕抖动，噗的一声，剑尖刺入了锦衫剑士的咽喉。

Experiment1: 基于LSTM的小说生成

青衣剑士凝视，十六名溜溜，你自己和？”范蠡微微一笑，又是秀丽那人的忠臣一个不在，我们吴吴吴国剑士身子伍子胥的奸臣，她逃不了薛先生，？我就，你这人，守者一道，我有高明去向王者又，各？？”阿青道“他他他他他他微微一笑，又放，遗命接连，以上拜访胡子，便挥剑她奏。吴王我羊儿间等我我我羊儿是你所是秀丽，他。

Experiment2: 基于Transformer的小说生成

第一回合双方僵持了一会儿后，锦衫剑士开大门进了他的房间，对方见他衣衫神色凝重，怕他进去不便脱身而进。他心中嘀咕不已，不由得长叹一声，只见

锦衫剑士左胸口猛地一扬，锦衫剑士竟伸缩右腿，直取敌人咽喉，当时已是深夜，这时长剑已脱

Conclusion

根据实验结果可知，在本次中文文本生成任务中，Transformer 模型展现出了显著优于 Seq2Seq 模型的性能，其生成的文本不仅连贯流畅，而且具备较强的逻辑性。而相比之下，Seq2Seq 模型生成的文本中则常常出现重复和无意义内容，整体质量较差。这一现象表明，尽管 Seq2Seq 模型结构较为简单，易于实现，但在处理复杂的自然语言生成任务时，其效果明显不如结构更为先进的 Transformer 模型。

具体来说，Seq2Seq 模型的优势在于能够在一定程度上保持句子的语法结构，生成的句子在形式上较为完整，虽然其中存在大量重复。然而，它的缺点也较为明显：首先是内容重复问题严重，生成的文本中常常出现重复词汇和短语，说明模型在生成过程中容易陷入循环；其次是缺乏上下文的连贯性，导致生成结果难以理解和实际应用；此外，Seq2Seq 在捕捉长距离依赖关系方面能力有限，进一步影响了生成文本的整体一致性和质量。

相较之下，Transformer 模型在文本生成方面具有明显优势。首先，它生成的文本在语义上更为通顺和连贯，能够有效避免重复和无意义的内容，体现出较强的语言建模能力；其次，Transformer 所采用的自注意力机制支持并行计算，大幅提升了训练和推理的效率；此外，该机制还能捕捉输入序列中的长距离依赖关系，使得模型在处理上下文信息时更加准确和灵活。然而，Transformer 模型也存在一定的不足，如其结构相对复杂，涉及多头注意力机制、位置编码等，需要更深入的理解和实现技巧；同时，其对计算资源的需求较高，在训练和部署时对硬件配置有一定要求。

综上所述，Transformer 模型虽然在实现上更为复杂，但在中文文本生成任务中凭借其优越的结构设计展现出了更强的性能，是应对复杂语言建模任务的更优选择。