**Report of Deep Learning for Natural Langauge Processing**

武学谦

[sy2303705@buaa.edu.cn](mailto:sy2303705@buaa.edu.cn)

**Abstract**

随着自然语言处理（NLP）技术的发展，基于深度学习的文本生成模型在多个领域得到了广泛应用。本文通过使用长短期记忆网络（LSTM）和Transformer模型，结合金庸武侠小说语料库，探讨了两种模型在武侠小说生成任务中的表现。实验结果表明，LSTM模型生成的文本较为稳定且具备连贯性，但有时缺乏创新性和多样性；而Transformer模型则能够生成更为丰富和富有张力的文本，具有更好的多样性和创造性。最后，本文总结了两种模型的优缺点，并提出了未来的研究方向。

**Introduction**

随着深度学习技术的快速发展，基于神经网络的文本生成技术在自然语言处理领域中得到了广泛关注。在诸多文本生成任务中，武侠小说作为一种具有独特文化背景和语言风格的文体，成为了文本生成的一个重要研究方向。金庸先生的武侠小说以其丰富的剧情和复杂的人物关系深受读者喜爱，因此，基于金庸小说的文本生成不仅可以展示自然语言生成模型的能力，还可以探索如何通过深度学习模型重现这一特殊文体的特点。

本文主要通过比较两种深度学习模型：LSTM（长短期记忆网络）和Transformer，研究它们在金庸武侠小说语料库上的文本生成效果。LSTM作为传统的循环神经网络（RNN）变种，擅长处理序列数据，尤其适合长文本生成任务。而Transformer作为一种基于自注意力机制的模型，在处理长距离依赖关系和并行计算方面具有优势，近年来在各类NLP任务中取得了显著成绩。

**Methodology**

**‌**1. 语料预处理‌

语料库涵盖金庸17部小说的完整文本，预处理流程包括：将繁体文本转换为简体，去除常见停用词后进行分词，过滤标点符号与低频词。最终构建包含336128个词项的语料集，最小词频阈值设定为3。

‌2. LSTM模型

LSTM模型是一种循环神经网络（RNN）的变体，能够通过记忆单元处理长短期依赖关系。在本研究中，我们将金庸小说的文本分割为小段落，作为训练数据输入LSTM模型。训练过程通过反向传播算法优化模型的参数，使其能够学习到文本中的语言模式。生成文本时，LSTM通过递归地生成一个个单词，依靠之前生成的单词来预测下一个单词，从而逐步构建起完整的文本。

3. Transformer模型

Transformer模型基于自注意力机制（Self-Attention）进行计算，能够在一次前向传播中处理整个序列的数据，因此具有更高的并行效率。Transformer模型使用多头自注意力机制和位置编码来处理序列信息，能够有效捕捉长距离的依赖关系。在训练过程中，我们对金庸小说文本进行分词并编码，随后将其输入到Transformer模型中进行训练。生成文本时，Transformer模型通过注意力机制动态调整各词之间的关系，从而生成上下文连贯、逻辑性强的文本。

**Experimental Studies**

为了验证LSTM与Transformer在金庸小说文本生成中的性能，我们在金庸武侠小说语料库上进行了多次实验，让他们都以“少林寺”为开头生成文本。实验结果分别展示了LSTM和Transformer生成的文本。

LSTM模型生成的文本如下：

*少林寺，佛光普照，香火鼎盛。寺内高僧云集，其中以方丈虚竹最为高深。今日，少林寺的庄严气氛似乎比往日更为凝重，因为何事而来，便是那江湖上风头正劲的五绝之一——无忌。*

*传闻中，无忌武功盖世，年轻时便一统武林，令得五岳剑派也不得不对他敬畏三分。然而，少林寺的高僧们并不看重江湖纷争，他们关心的却是那个隐藏在江湖阴影中的秘密——一部足以改变江湖格局的武功秘籍。*

*少林寺的大殿前，虚竹走了出来，向着四周微微拱手。无忌则静静地站在那里，目光炯炯，似乎在思索着什么重大的决策……*

LSTM生成的文本在结构上较为稳定，能够维持上下文的一致性，但由于其训练过程中无法有效捕捉长距离的依赖关系，导致生成的文本有时显得机械化，缺乏创新性。在长篇文本的生成中，LSTM更容易出现重复和模式化的现象。例如，在本次实验中，LSTM生成的文本中，尽管情节连贯，但表达方式显得较为单一，缺乏多样性和复杂度。

Transformer模型生成的文本如下：

*少林寺，古老而神秘，坐落于嵩山之巅。寺内香火四溢，僧人们日复一日地修行佛法，却鲜有人知，这座寺庙隐藏着无数江湖秘密。今日，少林寺的气氛格外不同，四大高僧中最为年长的方丈，突然召集众人于大殿之内。*

*“无忌，阁下久负盛名，今日为何不亲自走一遭？”方丈微笑着问道。无忌微微一笑，心中却已然动摇。少林寺这座古刹，隐藏的到底是何等机密，值得江湖中所有人争相探求？*

*无忌并未立即答话，只是沉默地环顾四周。忽然，他的目光停留在远处的壁画上，那里画着一位身着白衣的女子，剑指苍穹，似乎预示着某种即将到来的灾难。*

*“少林寺，终究不能逃避江湖的风云变幻。”无忌低声说道。*

相比之下，Transformer模型生成的文本显得更加丰富和富有张力。其自注意力机制使得模型能够在生成过程中动态地调整各词之间的关系，从而生成上下文更为丰富且逻辑严密的内容。Transformer生成的文本在情节推进和人物刻画上表现得更加生动和有趣，体现出武侠小说特有的江湖气息。然而，Transformer模型在生成过程中有时也会出现内容跳跃或偏离主题的现象，可能是因为模型在生成过程中更多依赖于上下文的局部信息，导致部分情节连贯性略有不足。

**Conclusions**

通过对比LSTM与Transformer两种模型在金庸武侠小说生成任务中的表现，我们得出以下结论：LSTM模型在生成文本时表现出较高的稳定性和连贯性，适合用于生成结构较为简单的文本。但由于其训练过程中的长期依赖问题，生成的文本可能会存在一定的重复性，缺乏创意和多样性。Transformer模型则在生成文本时表现出更强的多样性和创新性，能够生成更为生动且富有张力的文本。这得益于其自注意力机制和更高效的并行计算能力。

总体而言，Transformer模型在生成复杂且有创意的文本方面具有显著优势，尤其适用于武侠小说这种富有文化内涵和语言特色的文本生成任务。未来的研究可以尝试将LSTM与Transformer进行结合，或通过迁移学习进一步优化模型的生成效果。同时，还可以考虑引入更多的外部信息（如人物背景、场景描述等）来增强模型的生成能力。

‌