深度学习在自然语言处理中的应用研究

1. 摘要

本研究基于金庸武侠小说语料库，对比分析了两种不同架构的语言模型（LSTM与Transformer）的文本生成能力。实验采用全量训练策略训练LSTM模型，并对预训练的Transformer模型进行全参数微调。结果表明，当前基于Transformer的预训练语言模型在文本生成任务中表现优异，仅需少量微调即可生成风格鲜明的文本，而传统LSTM模型在长序列建模和生成质量上存在明显局限性。

2. 引言

近年来，深度学习技术的快速发展推动了自然语言生成（NLG）领域的进步，尤其在新闻撰写、对话系统、自动摘要及文学创作等场景中展现出巨大潜力。其中，针对特定文学风格（如武侠小说）的文本生成任务，成为兼具挑战性与应用价值的研究方向。

本研究选取LSTM和Transformer两类代表性模型进行对比实验。LSTM凭借其时序建模优势，曾是早期文本生成任务的主流选择；而Transformer凭借自注意力机制和并行计算能力，已成为当前自然语言处理领域的基准架构。通过分析两者在武侠小说生成任务中的表现，本研究旨在为风格化文本生成提供技术参考，并探讨不同模型的适用场景。

3. 模型架构

3.1. 1. LSTM模型

长短期记忆网络（LSTM）是RNN的改进变体，通过引入遗忘门、输入门和输出门机制，有效缓解了传统RNN的梯度消失问题。其核心公式如下：

- **遗忘门**:
  $$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$
- **输入门**:
  $i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$
  $\tilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$
- **记忆单元更新**:
  $C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$
- **输出门**:
  $o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$ $h_t = o_t \odot tanh(C_t)$

尽管LSTM能够捕捉局部上下文依赖，但其串行计算特性导致训练效率低下，且在生成长文本时易出现逻辑断层。

3.2. 2. Transformer模型

Transformer完全基于自注意力机制（Self-Attention），摒弃了循环结构，显著提升了长序列建模能力。其核心计算流程包括：

- **多头注意力**:
  $Attention(Q, K, V) = softmax(QK^T/\sqrt{d_k})V$

- **位置编码**：
  $PE_{(pos,2i)} = sin(pos/10000^{2i/d_model})$ $PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_model})$。

Transformer的并行化设计和全局依赖建模能力，使其在生成任务中表现出更强的鲁棒性和创造力。

---

4. 实验设计

4.1. 1. 数据准备

- **语料来源**：金庸武侠小说全集（如《射雕英雄传》《神雕侠侣》《笑傲江湖》等）。
- **预处理**：去除标点与特殊符号，按固定长度（100字符）切分序列，构建字符级输入。

4.2. 2. 训练配置

- **LSTM**：
  - 2层网络，隐藏层维度256
  - Adam优化器，学习率1e-3
  - 训练10轮
- **Transformer**：
  - 基于预训练模型GPT2（gpt2-chinese-cluecorpussmall）
  - 全参数微调5轮

4.3. 3. 生成结果对比

模型	输入示例	生成文本特点
LSTM	"在黄沙莽莽的回疆大漠之上"	局部连贯但逻辑跳跃，易出现重复短语（如"苏普""李文秀"频繁出现）。
GPT2微调前	"大明成祖皇帝永乐六年八月"	通用性强但风格偏离，生成内容多与历史事件杂糅（如提及"康熙""李自成"等无关人物）。
GPT2微调后	"一个嘶哑的嗓子低沉地叫着"	风格贴近武侠小说，情节连贯性显著提升（如生成"洪七公""计老人"等角色对话）。

---

5. 结论与展望

实验表明，Transformer模型在文本风格迁移、长序列建模和生成流畅性上均优于LSTM。预训练语言模型通过少量微调即可适配特定领域，展现出强大的泛化能力。未来工作可探索以下方向：

1. 结合更大规模的中文预训练模型（如文心一言）。
2. 采用参数高效微调技术（如LoRA）以降低计算成本。
3. 引入人工评估指标，定量分析生成文本的文学性。

本研究为风格化文本生成提供了实践案例，验证了Transformer架构在创造性任务中的优越性。