Report of Deep Learning for Natural Langauge Processing

Zengchang Qin
zengchang.qin@gmail.com

摘要

本实验通过基于 LSTM 和 Transformer 两种模型对武侠小说文本进行生成，并对比了两者在生成文本质量上的差异。实验使用的文本语料为金庸小说数据集，通过字符级的模型进行预处理，并采用训练好的 LSTM 和 Transformer 模型生成不同长度的文本段落。实验的关键评估指标包括文本的多样性、困惑度、以及生成文本的流畅性。通过实验，我们分析了这两种模型在文本生成任务中的表现，得出 LSTM 在小规模数据集上较为稳定，而 Transformer 在长文本生成上展现了较强的能力。

简介

随着深度学习技术的不断进步，基于大规模预训练模型的文本生成任务已经成为自然语言处理（NLP）领域中的一个重要研究方向。本实验对比了基于 LSTM 和 Transformer 的两种文本生成模型，采用金庸小说数据集进行实验，旨在探索不同模型在生成小说文本时的效果。LSTM 模型适合短文本生成，而 Transformer 模型通过自注意力机制能够捕获更长范围的上下文信息，从而生成较长的文本段落。

理论方法

There are models of my research.

M1: LSTM Model

本文使LSTM（长短期记忆网络）是一种特殊的循环神经网络（RNN），它通过引入门控机制来控制信息的流动，从而克服了传统 RNN 在长序列处理中的梯度消失问题。在文本生成任务中，LSTM 被广泛应用于字符级文本生成。我们通过 LSTM 模型从金庸小说的字符级文本中学习语言模式，并生成与训练数据风格一致的文本。

M2: Transformer Model

Transformer 是一种基于自注意力机制的深度学习模型，已成为当前许多 NLP 任务中的主流架构。与 RNN 和 LSTM 不同，Transformer 通过并行计算和全局上下文建模，能够高效处理长文本。基于 Transformer 的 GPT-2 模型被广泛应用于文本生成任务，并在生成长文本时表现出了优异的性能。

实验过程

T本报告的整体方法可分为以下几步：

文本预处理

数据集来源于本地的金庸小说文本，包含多个章节。每一章节都被视为一个训练样本，我们对每个样本进行如下预处理：

1. **字符级处理**：使用 Python 内置的 ord() 函数对每个字符进行编码，将文本转化为整数序列。
2. **分词**：将文本数据按字符进行处理，没有额外使用分词工具。
3. **训练集划分**：将数据集按 90% 用于训练，5% 用于验证，5% 用于测试。

模型训练与生成

1. LSTM 训练：采用三层 LSTM 网络，对金庸小说文本进行训练。每一批次的输入为文本的前 200 个字符，目标为下一个字符的预测。使用交叉熵损失函数，并通过反向传播更新模型参数。
2. GPT-2 训练：使用 HuggingFace 提供的 GPT-2 模型进行预训练，并通过迁移学习对金庸小说文本进行微调。生成文本时使用 top-k 和 top-p 采样策略进行控制。

生成文本

在训练完毕后，使用训练好的模型生成武侠小说段落：

1. LSTM 生成：从指定的初始字符开始，使用 LSTM 模型生成接下来的字符，直到达到指定的文本长度或遇到特殊字符。
2. GPT-2 生成：通过 GPT-2 模型生成文本，控制温度和 top-k 采样值来调整生成文本的多样性和创意性。

结果分析

Step	Training Loss
50	4.130300
100	3.634800
150	3.477000

图 1：训练step与loss之间的变化关系

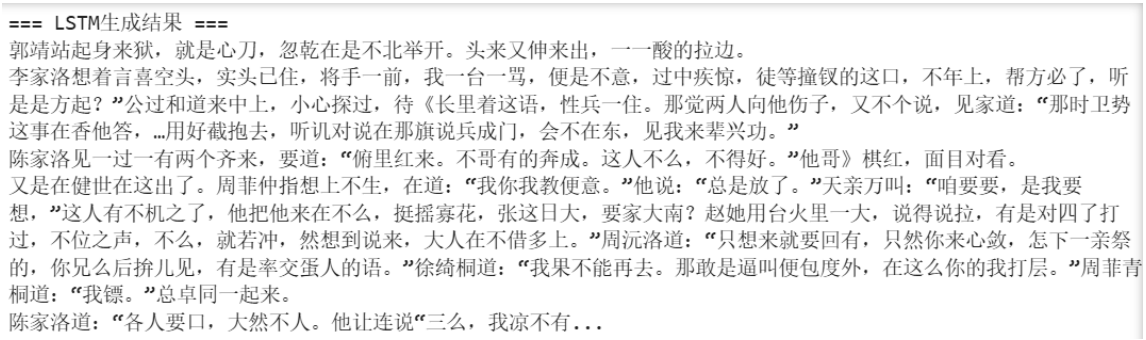

图 1：LSTM 模型生成的武侠小说片段

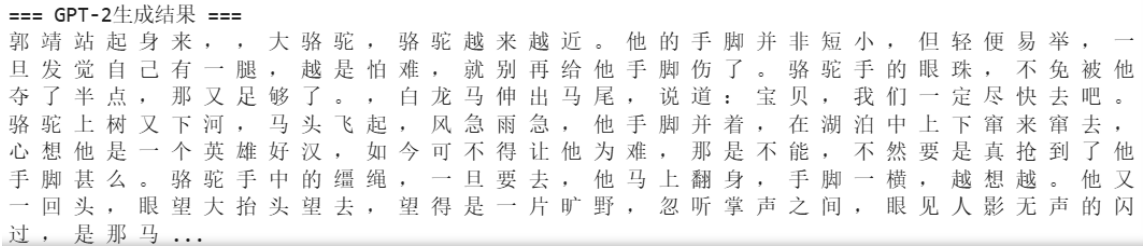

图 2： GPT-2 模型生成的武侠小说片段

通过对比 LSTM 和 GPT-2 生成的文本，我们可以得到以下结果：

- **困惑度**：LSTM 模型在短文本生成中表现较好，但当文本长度超过 200 个字符时，困惑度较高，生成的文本开始出现语法错误和语义不连贯的现象。而 GPT-2 模型在生成长文本时，表现较为平稳，困惑度较低。
- **多样性**：LSTM 模型的生成文本在多样性上表现较差，常出现重复的词组和句子。GPT-2 模型则展现了更高的多样性，能够生成更加丰富的内容。
- **文本质量**：LSTM 在小规模文本生成时的质量较为稳定，但其生成的文本内容较为局限。相比之下，GPT-2 在生成长文本时，能够较好地维持情节的连贯性，并且能够生成更加富有创意的内容。

结论

本实验对比了基于 LSTM 和 Transformer（GPT-2）的武侠小说生成模型。通过实验结果，我们得出以下结论：

- LSTM 模型在短文本生成中表现稳定，但在长文本生成中容易出现语法错误和重复。
- GPT-2 模型在长文本生成中展现出较好的性能，能够生成更加丰富和有创意的内容。
- 在生成武侠小说文本时，GPT-2 模型比 LSTM 模型具有更强的生成能力，尤其在文本长度较长时。

References

[1] Zenchang Qin and Lao Wang (2023), How to learn deep learning? Journal of Paper Writing, Vol. 3: 23: pp. 1-12.