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

何明璞

[**1113490780@qq.com**](mailto:1113490780@qq.com)

**Abstract**

本报告探讨了基于金庸小说语料库，分别使用长短期记忆网络（LSTM）和Transformer模型生成武侠小说的实验研究。我们介绍了这两种模型的训练方法，并对它们在生成文本时的连贯性、创造力和流畅性进行了比较。同时，我们分析了由于计算资源有限和训练数据不足，导致生成文本质量受限的原因。实验结果表明，尽管两种模型均能生成一定程度的小说化文本，但Transformer模型生成的文本更具连贯性和风格一致性。然而，由于训练数据量过少以及计算资源的局限，生成的文本质量仍有较大提升空间。

**Introduction**

武侠小说是中国文学中极具特色的一个类别，深受大众喜爱。作为一种文学形式，武侠小说不仅仅讲述了英雄人物的传奇故事，还承载了复杂的文化价值观和社会观念。金庸的武侠小说更是被誉为“现代武侠小说的巅峰”，其深刻的社会描写、复杂的情节设计、鲜明的人物塑造及独特的语言风格，使得金庸的作品成为研究和模仿的宝贵资料。随着人工智能技术的发展，特别是自然语言处理（NLP）领域的突破，利用机器学习模型自动生成武侠小说成为了一个富有挑战且有趣的研究方向。

在生成武侠小说的任务中，最重要的挑战之一是如何准确捕捉到金庸小说中复杂的语言风格与情节结构。这不仅要求模型能够生成流畅的语言，还需要模型具备一定的创意，能够在复杂的情境中产生合理的叙述和情节推进。为了达到这一目标，近年来出现了多种生成文本的深度学习模型，其中最为常见的两种模型是长短期记忆网络（LSTM）和Transformer模型。

LSTM是一种经典的递归神经网络（RNN），其通过引入“记忆单元”来解决传统RNN模型在长序列数据处理时遇到的梯度消失和爆炸问题。LSTM能够有效捕捉文本中的时间序列信息，尤其适合生成与前文紧密相关的文本。虽然LSTM在许多任务中表现良好，但其处理长距离依赖关系的能力仍然存在局限。

与LSTM相比，Transformer是一种基于自注意力机制的模型，它通过并行计算和全局依赖建模解决了LSTM在长序列文本生成中的缺陷。Transformer模型在捕捉长程依赖、并行计算和训练速度上都有显著优势，已广泛应用于机器翻译、文本生成等任务。然而，Transformer模型也存在一些挑战，如对计算资源的需求较高，且在小数据集上的训练效果往往不如预期。

本报告旨在探讨LSTM和Transformer模型在生成武侠小说文本中的应用，并对比它们在此任务中的优缺点。通过分析模型的训练过程和生成结果，我们希望能够总结出适用于武侠小说生成的模型架构，并为今后的研究提供参考。

**Methodology**

**‌**

为了进行这项实验，首先准备了金庸小说语料库，并对该语料库进行预处理，以适应LSTM和Transformer两种模型的训练要求。我们主要关注文本的词汇、句法结构和风格特征，以期能够生成与金庸作品风格接近的武侠小说。具体来说，我们的研究方法包含以下几个步骤：数据预处理、模型构建与训练、文本生成及评价。

1. 语料预处理‌

在数据预处理阶段，我们对金庸小说语料库进行了分词处理，并使用词嵌入技术将词语转换为向量表示。对于LSTM和Transformer两种模型，我们都使用了基于字符的处理方式，这样可以避免语言中的歧义和不同词语间的差异对模型性能的影响。通过这种方式，我们确保了模型可以学习到文本中的每个字符及其对应的上下文关系。

此外，为了提升训练效果，我们还进行了数据增强，利用金庸小说中的不同章节、场景和人物对训练数据进行了扩展。这样做的目的是让模型能够接触到更加丰富和多样化的情节内容，提高生成文本的多样性和创造力。

‌2. LSTM模型

长短期记忆网络（LSTM）是一种改进的递归神经网络（RNN），旨在解决传统RNN在处理长序列时存在的梯度消失和爆炸问题。LSTM通过引入“记忆单元”和三个门控机制（输入门、遗忘门和输出门），能够有效地捕捉和存储长时间跨度内的依赖信息，因此特别适合用于文本生成等需要捕捉上下文依赖关系的任务。

在本实验中，我们设计了一个标准的多层LSTM模型。模型的输入为金庸小说语料库中的字符序列，每个字符通过嵌入层（Embedding Layer）转化为一个稠密的向量表示。接着，这些向量进入LSTM层，通过LSTM单元逐步处理输入序列中的每一个字符。LSTM层能够保留来自前文的信息，并将其用于生成当前时刻的字符预测。

LSTM的核心优势在于其能够有效捕捉序列数据中的时序关系，尤其是短期和中期依赖。在武侠小说生成任务中，我们需要模型理解人物对话、场景描述以及情节的展开，这要求模型在生成每一个新字符时能基于前文信息进行合理预测。为此，我们在LSTM模型中使用了多层LSTM结构，这样可以让模型具备更强的表达能力，处理更复杂的语言结构。

3. Transformer模型

Transformer模型由Vaswani等人于2017年提出，采用了自注意力机制（Self-Attention Mechanism），并完全摒弃了传统的递归结构。Transformer通过并行计算和全局依赖建模，解决了LSTM在处理长序列时的瓶颈。与LSTM不同，Transformer模型能够同时处理输入序列中的所有元素，避免了LSTM的逐步计算和信息传递的局限。

Transformer模型由编码器（Encoder）和解码器（Decoder）两部分组成。在生成任务中，我们使用了一个简化的架构，仅使用了Transformer的解码器部分。模型的输入同样是金庸小说中的字符序列，首先通过嵌入层将每个字符转换为向量表示。然后，经过多个自注意力层和前馈神经网络层，模型生成每个字符的预测。

在自注意力机制中，模型能够为每个输入字符分配不同的权重，基于其他字符的信息动态调整自己的表示。具体来说，Transformer模型通过计算每个字符的“查询”（Query）、“键”（Key）和“值”（Value）来实现自注意力机制，通过这种方式，模型能够灵活地从全局范围内捕捉不同字符之间的关系，尤其是在处理长程依赖时表现出色。

**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：

*少林寺风刀铃声，云竹鸡飞寺门。佛像白蛇梦，香火没有。马蹄声踏月光，水池消失佛珠飞。猫影屋顶石，书卷忽然空。鹰鸟白云茶壶碎，和尚不动香案有风。门关了不见剑气，空地涌入墙壁轻。竹林不见人风，刀剑闭眼。水池开始飞去一只黄泥，风传声音月光山不见。墙裂地开，铃声止。豆腐跳下，香火没有。鹰白蛇转身，空中山外人已走。画裂了不再出现手，脚步轻如风大门突然关。少林寺月影斩，香案掉落一只茶。剑气飞过空中，眼不见一地步，风动不动不再开。人不见山门，白蛇回头空。脚步消失在地，不知何处风吹过。*

Transformer模型生成的文本如下：

*少林寺门一片云影白，豆腐跳过石壁，风轻如水。鹰飞茶壶掉下，香火不见人声。月光闪烁佛像动，地裂水涌猫上山。竹林风呼啸墙不动，和尚不动手不持。香案飞上屋顶，脚步轻如云，剑从池中出。少林寺的墙裂开，门不见白蛇飞。香火散去，月亮闭上眼。风轻如影，石头滚落。书卷失去颜色，白蛇从空中飞过，马蹄声不见。和尚坐下，空中一只鹰飞过。风从屋顶消失，脚步如水月无踪。一声钟响消失，风带走了茶，白蛇飞入佛像。竹林开口一片火，水池乱了，山顶天亮。门不见猫，豆腐飞向天空。脚步如影，佛像闭眼，少林寺风轻。*

*“少林寺，终究不能逃避江湖的风云变幻。”无忌低声说道。*

从实验结果可以看出，两种模型生成的文本既有一定的连贯性，但也存在着许多重复性或断裂的句子。LSTM模型较为频繁地出现重复的词语和句子，整体表现较为散乱；而Transformer模型生成的文本在流畅性和风格一致性方面较好，但仍然存在一些断裂和不连贯的图像描述。

可以得出以下几点结论：LSTM模型由于其顺序性处理的特点，难以有效捕捉长距离的依赖关系，导致生成的文本往往重复且不够连贯。而Transformer模型则通过自注意力机制，能够较好地捕捉长程依赖，使得生成的文本相对更加流畅和符合武侠小说的语言风格。然而，尽管Transformer模型在生成文本的连贯性和风格上有所优势，生成的文本仍然存在一些不连贯的地方，主要是因为训练数据量过少，导致模型未能完全掌握武侠小说的深层语言特征。

此外，由于实验是在计算资源有限的笔记本电脑上进行的，训练速度较慢，模型未能充分学习到金庸小说中的复杂结构和细节。因此，有限的计算资源和训练数据的不足对实验结果产生了显著影响。若能够使用更强大的计算设备并且增加数据量，模型的表现可能会有较大提升。

**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和Transformer模型在生成武侠小说方面各有优缺点。LSTM模型虽然能够生成连贯的句子，但在捕捉长程依赖和保持文本一致性方面存在困难；Transformer模型则能更好地捕捉文本中的长程依赖，生成更加流畅且符合武侠风格的文本，但仍然存在一定的生成不连贯的情况。由于训练数据不足和计算资源限制，生成文本的质量未能达到理想水平。未来的研究可以在更强的计算资源和更大规模的数据集上进行训练，以提高生成文本的质量。‌