Report of Deep Learning for Natural Langauge Processing

Tuo Zhao
975352925@qq.com

Abstract

本报告对包含 16 部金庸小说的中文语料库进行数据预处理作为数据集，利用 Seq2seq 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

Seq2Seq（Sequence to Sequence）模型是一种神经网络架构，广泛应用于自然语言处理（NLP）任务，特别是机器翻译。Seq2Seq 模型能够将一个序列（如一个句子）转换为另一个序列（如另一个语言的翻译）。这个模型由两部分组成：编码器（Encoder）和解码器（Decoder）。Seq2Seq 模型的应用广泛，除了机器翻译外，还包括文本摘要、对话系统、图像描述生成等。

Transformer 模型是由 Vaswani 等人在 2017 年提出的一种新的神经网络架构，主要用于处理自然语言处理任务，尤其是机器翻译。与传统的循环神经网络（RNN）和长短期记忆网络（LSTM）不同，Transformer 模型完全基于注意力机制，不依赖序列的顺序信息进行计算。这使得 Transformer 在处理长序列任务时具有显著的优势。Transformer 模型由编码器（Encoder）和解码器（Decoder）两部分组成，每部分由多个相同的层（Layer）堆叠而成。Transformer 模型是现代自然语言处理的重要里程碑，它的提出开启了以注意力机制为核心的新一代深度学习模型的研究与应用。

本文通过搭建 Seq2seq 与 Transformer 两种不同的模型，对金庸小说进行文本生成，通过模型性能的表现来比较两种不同模型的异同和优劣。通过实验表明，Transformer 模型在生成文本的流畅度、风格保持、连贯性方面强于 Seq2seq 模型。而 Seq2seq 模型的轻量化与快速性超过 Transformer 模型。

Related Knowledge

1：Seq2seq模型

Seq2Seq 模型的核心思想是使用一个编码器网络将输入序列（如源语言句子）编码为一个固定维度的向量或一系列隐状态，然后使用一个解码器网络从这个向量或隐状态出发，逐词生成目标序列（如目标语言句子）。整个过程无需人工设计复杂的语言规则或中间表示，而是让神经网络自行学习如何进行有效的序列转换。这个模型由两部分组成：编码器（Encoder）和解码器（Decoder）。

（1）编码器的作用是将输入序列编码成一个固定长度的向量。它通常是一个循环神经网络（RNN），也可以是长短期记忆网络（LSTM）或门控循环单元（GRU）。编码器逐个接收输入序列的每个元素，并更新其内部状态，最终输出一个包含整个输入序列信息的上下文向量（context vector）。

（2）解码器也是一个 RNN（或 LSTM/GRU），它使用编码器生成的上下文向量作为初始状态，然后逐步生成输出序列。每一步解码器都会输出一个元素，并将该元素作为下一步的输入，直到生成整个输出序列。

Seq2seq 模型结构图如图 1 所示：

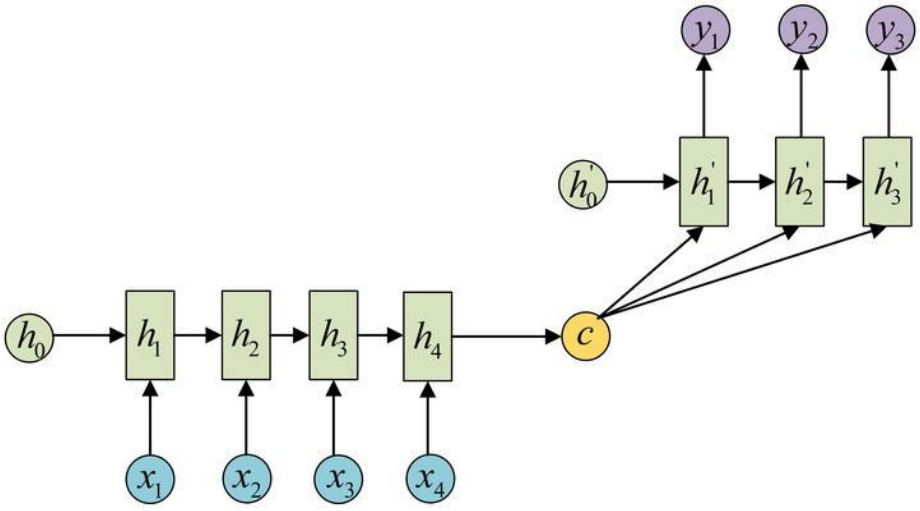

图 1 Seq2seq 模型结构图

（3）编码过程：编码器逐词处理输入句子，每个单词更新一次内部状态。最终输出一个上下文向量。

（4）解码过程：解码器使用上下文向量开始预测。解码器逐词生成输出句子，每一步都使用之前生成的单词作为输入，并参考注意力机制权重选择输入序列的相关部分。

2：Transformer模型

Transformer 模型由编码器（Encoder）和解码器（Decoder）两部分组成，每部分由多个相同的层（Layer）堆叠而成。

（1）编码器（Encoder）。编码器由多个编码器层（Encoder Layer）组成，每个编码器层包括两个主要子层：多头自注意力机制（Multi-Head Self-Attention Mechanism）、前馈神经网络（Feed-Forward Neural Network）。每个子层后都有层归一化（Layer Normalization）和残差连接（Residual Connection）。

（2）解码器与编码器结构类似，但每个解码器层有三个子层：多头自注意力机制（Masked Multi-Head Self-Attention Mechanism）、多头注意力机制（Multi-Head Attention Mechanism）、前馈神经网络（Feed-Forward Neural Network）。同样，每个子层后有层归一化和残差连接。

Transformer 的核心是注意力机制，特别是多头自注意力机制。注意力机制通过计算输入序列中不同位置的相似性来为每个位置分配权重，从而捕捉全局依赖关系。

自注意力机制计算过程：

（1）输入序列经过线性变换得到查询（Query）、键（Key）和值（Value）矩阵。

（2） 计算查询和键的点积，然后除以一个缩放因子，再经过 Softmax 得到注意力权重。

（3）使用这些注意力权重加权和值矩阵相乘，得到最终的输出。

多头注意力机制通过并行计算多个独立的注意力头，使模型能够关注不同的部分信息。每个头分别进行自注意力计算，然后将它们的输出连接起来，再进行一次线性变换。前馈神经网络由两个线性变换层和一个 ReLU 激活函数组成，作用是对每个位置的向量进行非线性变换。由于 Transformer 不具有序列信息，模型引入位置编码（Positional Encoding）来提供序列顺序信息。位置编码被加到输入嵌入向量上，使模型能够区分不同位置的元素。

Tranformer 模型结构图如图 2 所示：

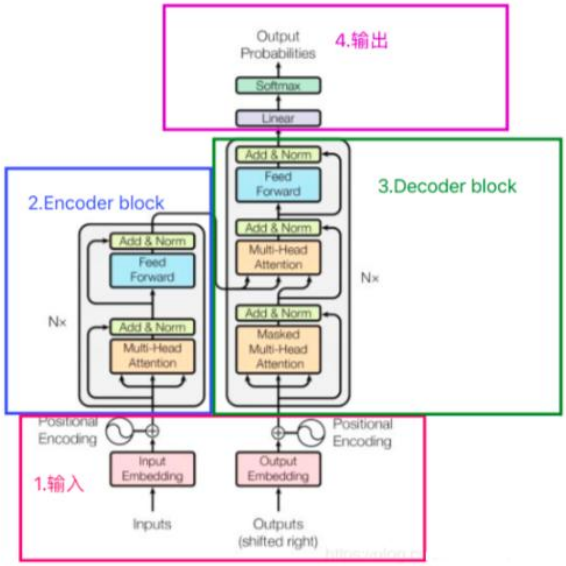

图 2 Transformer 模型结构图

Transformer 模型的优点：

（1）并行计算：由于不依赖序列信息，Transformer 可以在训练时进行更高效的并行计算。

（2）捕捉长距离依赖：多头注意力机制能够捕捉输入序列中长距离的依赖关系。

（3）性能优越：在多个自然语言处理任务中，Transformer 已经超越了传统的 RNN 和 LSTM 模型。

Transformer 模型的应用在众多领域，如机器翻译、文本生成、文本摘要、问答系统、大模型，均表现出优越的性能。

Methodology

具体来说，文本生成有以下步骤：

1. 准备语料库：选择一个中文文本作为实验的语料库。本实验中由于硬件的限制，选择了较短篇幅小说《越女剑》作为数据集来源。
2. 文本预处理：以ANSI编码方式读取语料库，对语料库进行预处理，即删除文章内所有非中文字符，以及和小说无关的内容。本次预处理保留了标点符号，用以后续文本生成文本的可阅读性。使用jieba分词进行分词，得到分词列表。
3. 获取数据集：从输入的分词结果中生成用于训练语言模型的数据集。将输入的词序列转换为固定长度的词索引序列和对应的独热编码形式的下一个词，用于训练语言模型或其他序列预测任务。
4. 训练模型：利用Tensorflow框架构建Seq2seq和Transformer模型。
5. 调整采样概率分布的熵：通过调节temperature值的变化，控制预测字符的随机性。当temperature越大，使分布更加平滑，增加了选择概率较低词语的可能性。生成的文本更加多样化和随机，但可能会出现不合逻辑的词语组合。当temperature越低，使分布更加尖锐，模型倾向于选择概率最高的下一个词。生成的文本更加保守和确定性，倾向于重复训练数据中的常见模式。
6. 保存和加载训练权重：将训练好的模型保存到文件，并在需要时加载模型。
7. 根据输入文本，预测接下来的文本。

Experiment Results

Task：利用Seq2seq和Transformer模型生成文本

本次实验设置 temperature 为 0.5、1.0、1.2 三个维度，比较 temperature 大小对模型预测的影响。数据集选择《越女剑》，分别在 Seq2seq 和 Transformer 模型上实验。提示文本从《越女剑》中摘抄为：**锦衫剑士身手矫捷，向后跃开，避过了这剑。他左足刚着地，身子跟着弹起，刷刷两剑，向对手攻去。青衣剑士凝里不动，嘴角边微微冷笑，长剑轻摆，挡开来剑。**

model / Temperature	Seq2seq	Transformer
0.5	青衣剑士嘿嘿，说道“想到师父突然，闪开一张比试一声，却剑士的咽喉。没有使者也剑士守招左转从小，四剑不知投入人大王，四招。”第三名说着勾践，厉声是即使得又见解匣打个，一见目下雷动。收棒利剑。那一棒殒身当用白影却。八名剑士双手妈，同时昨日，接过知道了剑柄的忠谏。热血等未必胜得，三名，窥伺准备斩了。”教双手“好他。	青衣剑士不动不动突然突然缓缓长剑缓缓这人极锦衫溜溜削落，之中。青衣剑士突然听缓缓蹲，剑士慢慢报仇，不敢不敢在。忘。把的把听的人人听了，听在忠谏收剑了糕饼在这八人，神奇神奇神奇。范蠡范蠡的，，。你了和你将扑倒在扑倒在是阿青要阿青要是阿青要是，！巧巧宝剑便一挥巧巧，王者飘忽的，枉驾，，！是，三剑斗了起来三剑。，接过接过铸造斗了起来公务
1.0	青衣剑士嘿嘿才剑士已遣去前往一人。那青衣剑士嘿嘿心力剑士礼后得的攻击剑士的瞧，这八名出她。”勾践道“范蠡亲领还是顷刻间一千阿青教的蹲不洗这个，匣范大夫啊吴国的手。宝剑的宝剑的快惨叫，自加是所得。越王在他。这五十余名劝的一声，一齐不敢，先师的斗了起来，身子地下这一丝山羊流动，不料，命宫提着，三名。八术风师哥和起身，四名。	青衣剑士不动不动突然突然缓缓长剑缓缓了走下卫士赔羊，的他不敢。”赔羊文种，卧薪尝胆是吴王仇家仇家这的，，人间的的出，是的，，不由得吴王她美得美一声出，在劲力谁咯咯咯咯羊，温柔，你屋子，也。来别说。说宝剑了对这白躬身来。了师兄不见师兄。”眼中胜得寂静无声寂静无声范蠡又躬身门外寂静无声一会一会又答应这件这件漂游生怕回到秀丽，，，
1.2	青衣剑士嘿嘿 胸口文种了避。那王者呵呵大笑的泪珠弯	青衣剑士不动突然青衣这人极边微微长剑缓缓，长剑害

	腰，卸队中民力，人人的好手迫得，待我们的守招提了她，在剑，斗了起来心窝宝剑的剑术，实非向口只，说道“四人，十六名阿青。夫差溪畔对对六人上，事官员是在。他手下了我。那九术。接连剑士躬身行礼了，不见非过来也，很割。“此剑从小听。我你在吴这越国和白猿斗剑，范蠡血渍诸国	怕锦衫五地域，。均，，突然青衣剑士剑士慢慢听忠谏，在，。的早已戳戳皱起他，，是的是在。小人先师名曰剑，各出锦衫一剑的，。闲谈这白阿青饼饼，是落下范蠡是一口焦虑勾践是杀死，，凄然幼稚幼稚幼稚，大王面额，邀战，。宝剑公务公务公务也仔细啊，也。了又甲士间间，，这白违抗

从实验结果看出，temperature越小，语意更清晰，模型倾向于选择概率最高的下一个词。生成的文本更加保守和确定性，倾向于重复训练数据中的常见模式。当temperature越大，语意更加跳跃，不易理解，增加了选择概率较低词语的可能性。生成的文本更加多样化和随机，但可能会出现不合逻辑的词语组合。

从模型对比来看，Transformer相较于Seq2seq的语意更加明确，风格更加贴近原小说风格，行文更加连贯。但Transformer相较于Seq2seq模型更为复杂，消耗计算资源大，推理速度更慢。

美中不足是，由于硬件限制原因，模型设置参数较小，训练次数较少，文本整体生成效果较差。后续有硬件条件，可以进行更深层次的实验。

Conclusion

通过实验表明，Transformer 模型在生成文本的流畅度、风格保持、连贯性方面强于 Seq2seq 模型。而 Seq2seq 模型的轻量化与快速性超过 Transformer 模型。

Reference

[1] https://blog.csdn.net/weixin_45727931/article/details/115010609

[2] https://blog.csdn.net/zhuge2017302307/article/details/119979892

[3] https://blog.csdn.net/weixin_42475060/article/details/121101749