Report of Deep Learning for Natural Langauge Processing

Seq2seq 和 transformer 用于文本生成

韩子轩
18210237202@163.com

Abstract

本篇报告从金庸的小说中选取部分经典作品作为语料库，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），并对比与讨论了这两种方法的优缺点。

Introduction

利用给定金庸小说语料库，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。

Methodology

M1: Seq2Seq 模型

Seq2Seq（Sequence to Sequence）是一种能够完成输入时序序列向输出时序序列转化的深度学习模型，在自然语言处理领域应用空间较广。该模型的核心思想是使用两个 LSTM（长短期记忆网络）模块，分别作为模型的编码器（Encoder）和解码器（Decoder），实现输入序列到输出序列的映射。

Seq2Seq 模型的结构由编码器、解码器和上下文向量组成，编码器接收输入序列和，通过 LSTM 网络对输入进行压缩，得到上下文向量。将上下文序列作为输入向量输入至解码器网络中，得到输出文本。解码器可以根据实际需求换用 RNN、LSTM 和 GRU 网络。

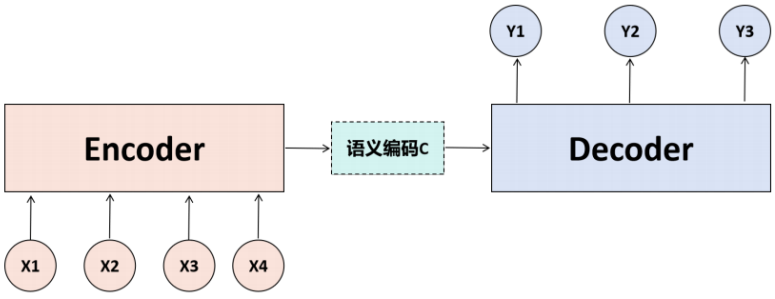

图 1 Seq2Seq 结构图

M2: Transformer 模型

Transformer 模型是一种基于注意力机制的深度学习模型架构，由 Vaswani 等人在 2017 年提出，主要用于自然语言处理任务。与传统的 Seq2Seq 模型不同，Transformer 完全抛弃了循环神经网络（RNN），依赖自注意力机制（Self-Attention）来处理序列数据，极大地提升了并行处理能力和模型效率。Transformer 模型主要由编码器（Encoder）和解码器（Decoder）两部分组成，每部分都由多个层（Layer）堆叠而成。

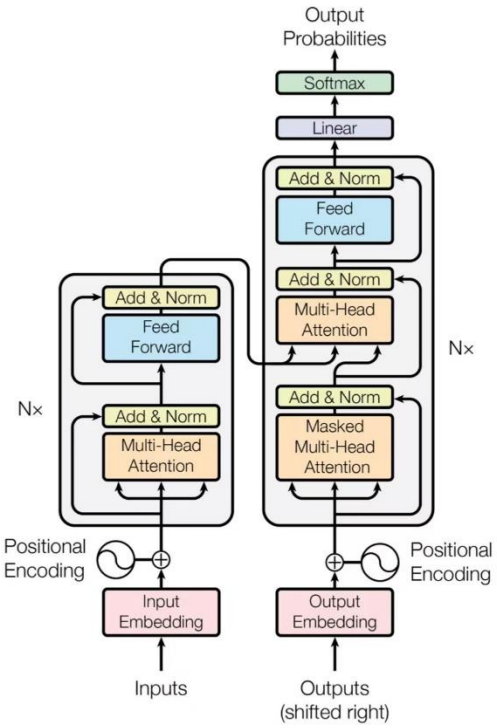

图 2 Transformer 结构图

(1) 编码器（Encoder）

编码器由多个相同的层（Layer）堆叠而成，每个层包括两个子层：

1. 多头自注意力机制（Multi-Head Self-Attention）

2. 前馈神经网络（Feed-Forward Neural Network）

多头自注意力机制（Multi-Head Self-Attention）自注意力机制允许每个位置的表示根据序列

中所有其他位置的表示进行加权和。多头注意力机制将输入分成多个头，每个头独立地执行注意力计算，然后将结果拼接并线性变换。

$$Attention(Q,K,V) = softmax(\frac{QKT}{\sqrt{d_k}}V)$$

其中，$Q(Query)$、$K(key)$ 和 V $(Value)$ 是通过输入序列线性变换得到的。

多头注意力机制可以表示为：

$$MultiHead(Q,K,V) = Concat(head1,\ldots,head2)Wo$$

其中，$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$

前馈神经网络（Feed-Forward Neural Network）前馈神经网络包括两个线性变换和一个激活函数，通常使用 ReLU：

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2$$

(2) 解码器（Decoder）

解码器的结构与编码器类似，但多了一个编码器-解码器注意力机制层（Encoder-Decoder Attention），用于将编码器的输出信息引入解码过程。

每个解码器层包括三个子层：

1. 多头自注意力机制（Multi-Head Self-Attention）
2. 编码器-解码器注意力机制（Encoder-Decoder Attention）
3. 前馈神经网络（Feed-Forward Neural Network）

(3) 位置编码（Positional Encoding）

由于 Transformer 模型没有循环或卷积结构，因此引入位置编码（Positional Encoding）来

注入序列的位置信息。位置编码可以通过正弦和余弦函数计算得到：

$$PE_{pos,2i} = sin(pos/10000^{2i/dmodel})$$

$$PE_{pos,2i+1} = cos(pos/10000^{2i/dmodel})$$

其中，pos 是位置，i 是维度索引，d_{model} 是模型维度。

Transformer 模型通过自注意力机制和并行计算的特点，极大地提升了处理长序列的能力和计算效率。Transformer 不仅在机器翻译等自然语言处理任务中取得了显著的成功，还被广泛应用于图像处理、语音识别等领域。特别是其变体 BERT 和 GPT 在各种任务上的表现，使其成为现代深度学习的重要工具。

Experimental Studies

1. 训练文本与测试样本生成

- **数据预处理：**从给定文本文件中删除所有特殊字符（如数字、英文字母等），然后以句号“。”为界限，将整个文本分割成单独的句子。
- **筛选标准：**

    - 句子中必须含有字符“她”。
    - 句子长度应在 10 到 40 个字符之间。
    - 该句的下一句话长度也应在 10 到 40 个字符之间。
- **样本选择**：筛选出 300 句符合上述条件的句子作为训练样本，每个样本的训练目标是其后的一句话。同时额外选取 10 句符合条件且与训练样本不重复的句子作为测试样本。

2. 模型构建

- Transformer 模型：
    - 编码器和解码器的嵌入层维度均设置为 256。
    - 隐藏层维度均设置为 512。
    - 使用了 8 个注意力头和 2 层编码器、解码器层。
    - 使用了位置编码和 dropout。
- Seq2Seq 模型：
    - 编码器和解码器均采用 LSTM 模型。
    - 编码器和解码器的嵌入维度均设为 150。
    - 编码器和解码器的隐藏层维度均设为 100。

3. One-hot 字典生成

- **字符编码**：为处理得到的训练样本和测试样本中的每一个字符进行不重复地编号，从而构建一个 one-hot 编码字典。
- **特殊标识符**：字典中包括特殊标识符“<PAD>”、 “<BOS>” 和“<EOS>”。

4. 批次数据对齐处理

- 对齐方式：
    - 每个序列的开始添加一个特殊的开始标识符“<BOS>”。
    - 在序列末尾添加结束标识符“<EOS>”。
    - 在序列末尾添加“<PAD>”填充标识符，直到达到该批次中最长序列的长度。

5. 模型训练设置

- 训练参数：
    - 迭代训练 50 代。
    - 批次大小设置为 2。
    - 学习率设置为 0.001。
- **优化器**：使用 Adam 优化器。
- **损失函数**：使用交叉熵损失函数

6. 训练和结果展示

- **模型训练**：循环迭代训练模型，每个 epoch 结束后打印损失值。
- **结果展示**：训练完成后，生成测试样本的目标句子并展示源句子、真实目标句子和生成的目标句子。

2. 实验结果

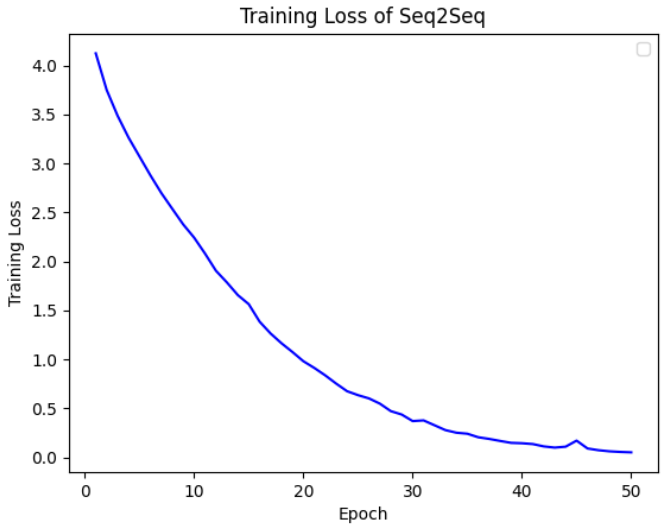

图 3 Seq2Seq 损失函数收敛过程

表 1 Seq2Seq 文本生成结果

小说原文	小说下一句	实际输出
段正淳见她不明世事，更是难过，说道：“婉儿，日后我要好好待你，方能补我一些过失	你有什么心愿，说给我听，我一定尽力给你办到	”镇南王陪笑道：“我去牵马
”就此痴痴的目不转睛的凝视着她	王语嫣双颊晕红，转开了头，心想：“这人如此瞧我，好生无礼	”那少女缓步走到青石凳前，轻轻巧巧的坐了下来，却并不叫段誉也坐
他耳叫得阿紫随后跟来，当下加快脚步，几步跨出，便已将她抛得老远	忽听得阿紫娇声说道：“姊夫，姊夫，你等等我，我……我跟不上啦	前此后数月之中，除了大风雪，两人总是是在外漫游
”说着走过去扶她，手掌尚未碰到她手臂，突然间拍的一声，左颊上热辣辣的吃了一记耳光	她虽在重伤之余，出手仍是极为沉重	那宫人佯怒气，已钻到了马腹之下，飕飕飕连射三箭
这两人相距尚远，他凝神听去，辨出来者是两个女子，心道：“多半是阿紫和她妈妈来了	嗯，我要问明段夫人，这幅字是不是段正淳写的	”阿朱捏中一声，木姑娘最的心中，可是可是可以以解开，在外漫游

图 4 Transformer 损失函数收敛过程

表 2 Transformer 文本生成结果

小说原文	小说下一句	实际输出
众少年一听，都是十分兴奋，均想：“就算公主挑不中我，我总也亲眼见到了她	西夏人都说他们公主千娇百媚，容貌天下无双，总须见上一见，也不枉了远道跋涉一场	不的木
虚竹大吃一惊：“糟糕，她摸到了我的光头	”岂知那少女所摸到的却是一片短发	
钟姑娘苦待救援，渡日如年，她如见我既不回去，她父亲又不来相救，只道我没给她送信	好歹我得赶到无量山去，和她死在一块，也好教她明白我决不相负之意	誉你我只，
他左手抓住了阿紫背心，右手点了她穴道，柔丝索早已缩入了大袖之中	他掷尸、挥索、招手、擒人，一直在哈哈大笑，待将阿紫擒到手中，笑声仍未断绝	,，这定
萧峰拾起断刀，一一拍落，跟着手一挥，那断刀倒飞出去，拍的一声，刀柄撞在她腰间	那年轻女子“啊”的一声叫，穴道正被撞中，身子也登时给定住了	得，了，“,，的

Conclusions

从训练结果可以看出，Seq2seq 模型损失函数在 50 轮次的训练内迅速收敛，生成的文本语句通顺，文本语意较为完整，尽管与原句不完全一致，但与上一句仍存在一定的逻辑关系。

Transformer 模型下，损失值始终维持在较高的区域，并没有明显的收敛趋势，而 Transformer 模型生成的文本基本都是一两个字的短词，几乎不成整句，毫无语意可言，更不用提与上一句是否存在逻辑关系。这是因为训练的数据量太小，而 Transformer 模型达到较好的性能表现通常需要经过较大的数据。为了应对这种情况，可以使用成熟的模型进行文本生成，如调用 gpt2API。