Report of Deep Learning for Natural Langauge Processing

BaobaoHe　ZY2354234
hebaobao@buaa.edu.cn

Abstract

本文利用给定语料库（金庸语小说语料库），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 。

Introduction

Seq2seq 是用于自然语言处理的一组机器学习方法。应用包括语言翻译、图像字幕、对话模型和文本摘要。 Seq2seq 使用序列转换：它将一个序列转换为另一个序列。

seq2seq 模型由一个编码器和一个解码器组成，这两个解码器通常以 RNN 的形式实现。编码器捕获输入序列的上下文并将其发送到解码器，然后解码器生成最终的输出序列。

编码器负责处理输入序列并捕获其基本信息，这些信息存储为网络的隐藏状态，并在具有注意机制的模型中存储为上下文向量。上下文向量是输入隐藏状态的加权和，针对输出序列中的每个时间实例生成。

解码器从编码器获取上下文向量和隐藏状态，并生成最终输出序列。解码器以自回归方式运行，每次生成输出序列的一个元素。在每个步骤中，它都会考虑先前生成的元素、上下文向量和输入序列信息，以预测输出序列中的下一个元素。具体而言，在具有注意力机制的模型中，上下文向量和隐藏状态连接在一起形成一个注意力隐藏向量，用作解码器的输入。

注意力机制是 Bahdanau 等人于 2014 年引入的一种增强机制，用于解决基本 Seq2Seq 架构的局限性，即较长的输入序列会导致编码器的隐藏状态输出与解码器无关。它使模型能够在解码过程中有选择地关注输入序列的不同部分。在每个解码器步骤中，对齐模型使用当前解码器状态和所有注意隐藏向量作为输入来计算注意力分数。对齐模型是另一种与 seq2seq 模型联合训练的神经网络模型，用于计算隐藏状态表示的输入与注意力隐藏状态表示的先前输出的匹配程度。然后将 softmax 函数应用于注意力分数以获得注意力权重。

在某些模型中，编码器状态直接输入到激活函数中，无需对齐模型。激活函数接收一个解码器状态和一个编码器状态，并返回其相关性的标量值。

Transformer 模型（直译为“变换器”）是一种采用注意力机制的深度学习模型，这一机制可以按输入数据各部分重要性的不同而分配不同的权重。该模型主要用于自然语言处理（NLP）与计算机视觉（CV）领域。

与循环神经网络（RNN）一样，Transformer 模型旨在处理自然语言等顺序输入数据，可应用于翻译、文本摘要等任务。而与 RNN 不同的是，Transformer 模型能够一次性处理所有输入数据。注意力机制可以为输入序列中的任意位置提供上下文。如果输入数据是自然语言，则 Transformer 不必像 RNN 一样一次只处理一个单词，这种架构允许更多的并行计算，并以此减少训练时间。

Transformer 模型于 2017 年由谷歌大脑的一个团队推出，现已逐步取代长短期记忆（LSTM）等 RNN 模型成为了 NLP 问题的首选模型。并行化优势允许其在更大的数据集上进行训练。这也促成了 BERT、GPT 等预训练模型的发展。这些系统使用了维基百科、Common Crawl 等大型语料库进行训练，并可以针对特定任务进行微调。

Experimental Studies

Seq2Seq 模型

在本实验中，构建一个基于 Seq2Seq（序列到序列）模型的句子生成系统，利用 Python 和 PyTorch 实现。目标是通过输入源语言句子，生成目标语言的对应输出。

一、模型构建

1. 数据集类（CorpusDataset）：

   初始化函数：加载源数据、目标数据及其词汇映射。

   `__getitem__`函数：获取指定索引的源句子和目标句子，并将其转换为索引。

   批量数据对齐函数：对源句子和目标句子进行填充对齐，并转换为 PyTorch 张量。

   `__len__`函数：返回数据集的长度。

2. 编码器类（Encoder）：

   嵌入层：将词汇转换为向量。

   LSTM 层：处理序列数据，生成隐藏状态。

3. 解码器类（Decoder）：

   嵌入层：将词汇转换为向量。

   LSTM 层：处理序列数据，从隐藏状态开始生成目标序列。

4. Seq2Seq 模型类（Seq2Seq）：

   初始化编码器和解码器。

   线性层：将解码器输出转换为词汇概率分布。

   交叉熵损失函数：计算预测结果与真实标签之间的损失。

二、训练过程

1. 参数设置：

   设置设备（GPU 或 CPU）、批量大小、语料数量、测试语料数量、学习率等。

   初始化数据集和数据加载器。

2. 模型训练：

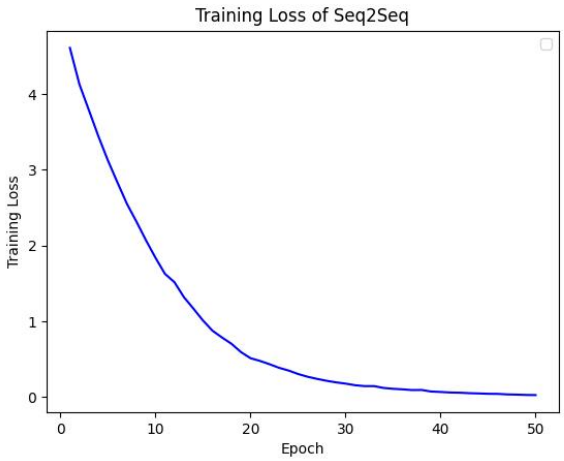
   初始化模型和优化器。

   进行多轮训练，在每轮中计算损失并进行反向传播以更新模型参数。

   记录并打印每轮的训练损失。

3. 绘制损失曲线：

   使用 Matplotlib 绘制训练损失随轮次变化的曲线，结果如下：



三、句子生成

1. 定义生成函数:

输入源句子，转换为索引并输入编码器。

使用解码器逐步生成目标句子，直到生成结束符或句子长度超过 40。

2. 模型评估:

将模型设置为评估模式。

遍历测试数据，生成目标句子并打印源句子、真实目标句子和生成的目标句子。

四、结果展示

保存测试数据中的源句子、真实目标句子和生成的目标句子，结果如下:

----------Result 1----------
Source sentence: 那女郎站起身来，不料段誉慌乱中兀是持着尸体，将死尸的脑袋向着她胸口撞去
True target sentence: 那女郎在死尸脑袋上一推，段誉“啊”的一声，摔了出去，尸体正好压在他身上
Generated target sentence: 原来只是她身子，身上下拜，总是不求人的欺负

----------Result 2----------
Source sentence: 至于石阶上和她胸口嘴边的鲜血，那是她预先备下的鸡血，原是要诱敌人上钩之用
True target sentence: 不料李秋水十分机警，明明见她已然断气，仍是再在她胸口印上一掌
Generated target sentence: ”徐老夫人道：“你的短箭见血封喉，剧毒无比

----------Result 3----------
Source sentence: 你那会种花的师妹躲哪里去了？我终究找得到她
True target sentence: 第六句你回答不医，我去杀了你那个美貌师妹
Generated target sentence: 你多半就，你觉好了呢

----------Result 4----------
Source sentence: 那是她出于爱我的一片痴心，手段虽然过份，总也不是歹意
True target sentence: ”言念及此，便即宽心
Generated target sentence: ”有一名帮众递过火把，司空玄拿在手里，走上两步

----------Result 5----------
Source sentence: 心想：“她父母回来，多半要挖开坟来看个究竟
True target sentence: 须得在墓前竖上块牌子才是
Generated target sentence: ”那知道姑娘，我真担心你好不可

----------Result 6----------
Source sentence: 但如果她说：“不用忙，我还有话跟你说
True target sentence: ’那么我便等着，瞧她有什么话吩咐
Generated target sentence: ”当下说道：“这枚七宝指环，你是从哪里偷来的？”语音严峻，如审盗贼

----------Result 7----------
Source sentence: 阿碧在她耳边道：“阿朱阿姊，赶走了敌人之后，我来帮你收作
True target sentence: ”阿朱捏了捏她的手示谢
Generated target sentence: ”阿紫扁了嘴起来：“有奸细！有刺客！”还不知道二人乃是萧峰和阿紫

----------Result 8----------
Source sentence: ”萧峰跌足道：“唉，小孩子知道什么？我要问她一件事
True target sentence: 这世上只有她一个人知道
Generated target sentence: ”童姥断腿几，好进来，好生无礼

----------Result 9----------
Source sentence: 因此她是有力偷袭，无力还手
True target sentence: 你如杀她，那便是改了你的规矩，你如改了规矩，那便是乌龟儿子王八蛋
Generated target sentence: 可如此刻伴着那山羊胡子司空玄，实在无味得紧

----------Result 10----------
Source sentence: 段誉抢过去挡在她身前，叫道：“你躲在我後面
True target sentence: ”便在这时，旭摩智双手已扣住他□喉，用力收紧
Generated target sentence: ”童姥断好，好，好生的一种生的惻隐

Transformer 模型

一、定义数据集类

CorpusDataset：处理源语言和目标语言数据，包含批量数据对齐功能。

二、定义模型相关类

PositionalEncoding：用于计算位置编码，帮助 Transformer 模型理解序列数据中的位置信息。

TransformerModel：包含嵌入层、位置编码层、Transformer 编码器和解码器层。

generate_square_subsequent_mask：生成上三角矩阵掩码，用于屏蔽未来时间步的信息。

forward：模型的前向传播，包括嵌入、位置编码、Transformer 层和解码层。

三、数据处理

1.生成句子函数

generate_sentence_transformer：使用 Transformer 模型生成句子。

输入：句子、模型、最大长度。

输出：生成的句子。

2.检查函数

check_vocab_construction：检查词汇表的构建过程。

check_dataset_indexing：检查数据集的索引转换。

check_dataloader：检查数据加载器。

check_special_tokens_application：检查特殊标记的应用。

四、主程序

1.设备选择：根据是否有 GPU 选择设备。

2.批处理大小和数据量：设置批处理大小为 2，语料库数量为 300，测试语料库数量为 10。

3.文件读取和预处理：从文本文件读取数据，处理并筛选符合条件的句子对。

4.构建词汇表：创建单词到索引和索引到单词的映射。

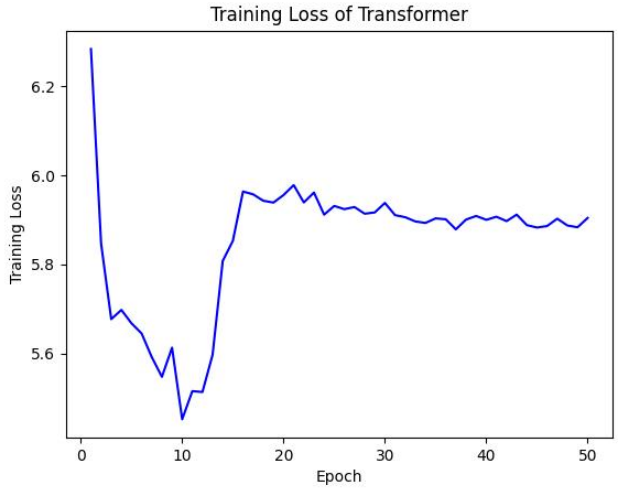
5.数据集和数据加载器：创建数据集和数据加载器。

6.模型初始化：初始化 Transformer 模型。

7.模型训练：设置优化器和训练参数，进行模型训练。

五、结果

损失函数与结果如下所示：



```
----------Result 1----------
Source sentence: 他双眼一瞬不瞬的瞧着阿朱，只要几把泥土一撒下去，那便是从此不能再见到她了
True target sentence: 耳中隐隐约约的似乎听到她的话声，约定到雁门关外骑马打猎、牧牛放羊，要陪他一辈子
Generated target sentence:
----------Result 2----------
Source sentence: 这世上只有她一个人知道
True target sentence: 若不是你来打岔，她已经说出来了
Generated target sentence: 休下任你
----------Result 3----------
Source sentence: ”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
True target sentence: 你有什么心愿，说给我听，我一定尽力给你办到
Generated target sentence:
----------Result 4----------
Source sentence: 群豪见她眼眶中鲜血流出，掠过她雪白的脸庞，人人心下几怖，见她走来，便都让开了惊步
True target sentence: 只见她笔直向前走去，渐渐走近山边的深谷
Generated target sentence: 的到：，‘，，
----------Result 5----------
Source sentence: 她与无崖子是同门师姊弟，一脉相传，武功的子完全一般
True target sentence: 虚竹依法修习，进展甚速
Generated target sentence: 这是声，“以道，，，得，年，不，
----------Result 6----------
Source sentence: 嘿嘿，终于是她先我而死
True target sentence: 她全身骨碎筋断，吐气散功，这样的死法，却是假装不来的
Generated target sentence: ，她
----------Result 7----------
Source sentence: 那深情关切之意，仍然留在她的眉梢嘴角
True target sentence: 萧峰大叫一声：“阿朱！”抱着她身子，向荒野中直奔
Generated target sentence: 我疑面了
----------Result 8----------
Source sentence: 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
True target sentence: 嗯，我要问明段夫人，这幅字是不是段正淳写的
Generated target sentence: 是这一，：，，，！，，个在”不那，了，，不她，，增夫女，身一：”有在
----------Result 9----------
Source sentence: 段誉道：“她……她是木姑娘，是儿子结交的……结交的好朋友
True target sentence: ”镇南王见了儿子神色，已知其意，见木婉清容颜秀丽，暗暗喝采：“誉儿眼光倒是不错
Generated target sentence: 的
----------Result 10----------
Source sentence: 那人道：“很好，我等在这里，你去请她指点杀我的法门
True target sentence: ”段誉道：“我不要杀你
Generated target sentence: 一这，，，
```

Conclusions

通过对金庸小说语料库的实验，在文本生成任务中，Seq2Seq（Sequence to Sequence）模型和 Transformer 模型是两种常用的方法。它们各有优缺点，下面将对这两种模型进行详细比较。

Seq2Seq 模型更适合处理较短的文本生成任务，具有较成熟的应用经验和较低的计算资源需求，但在处理长序列和复杂依赖关系时效果较差。

Transformer 模型在处理长序列和复杂依赖关系时表现优异，具备更强的表达能力和更快的训练速度，但对计算资源要求较高，且在小数据集上容易过拟合。

选择具体的模型应根据任务的具体需求、数据集的大小和计算资源等因素综合考虑。