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

Weida Chen

908715799@qq.com

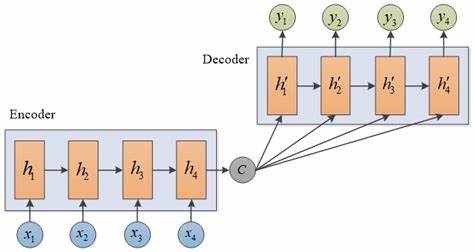
**Abstract**

本研究利用给定的金庸小说语料库，用Seq2Seq与Transformer两种不同的模型实现了文本生成的任务（给定开头后生成武侠小说的片段或者章节），并对比与讨论了两种方法的优缺点。

**Introduction**

1. Seq2Seq模型

Seq2Seq（Sequence to Sequence）模型是一种用于处理序列数据的深度学习模型，广泛应用于机器翻译、文本摘要、对话系统等自然语言处理任务。它的核心思想是将输入序列编码为一个固定长度的向量，再将该向量解码为目标序列。Seq2Seq 模型通常由编码器（Encoder）和解码器（Decoder）两部分组成。其结构如下图所示。



图一、Seq2Seq模型网络架构

下面分别介绍Seq2Seq模型中的编码器和解码器两大部分。

1. 编码器（Encoder）

编码器接收输入序列并将其转换为一个固定长度的向量。编码器通常由递归神经网络（RNN）、长短期记忆网络（LSTM）或门控循环单元（GRU）构成。输入序列的每个元素依次输入编码器，最后一步的隐藏状态被认为是整个输入序列的表示。

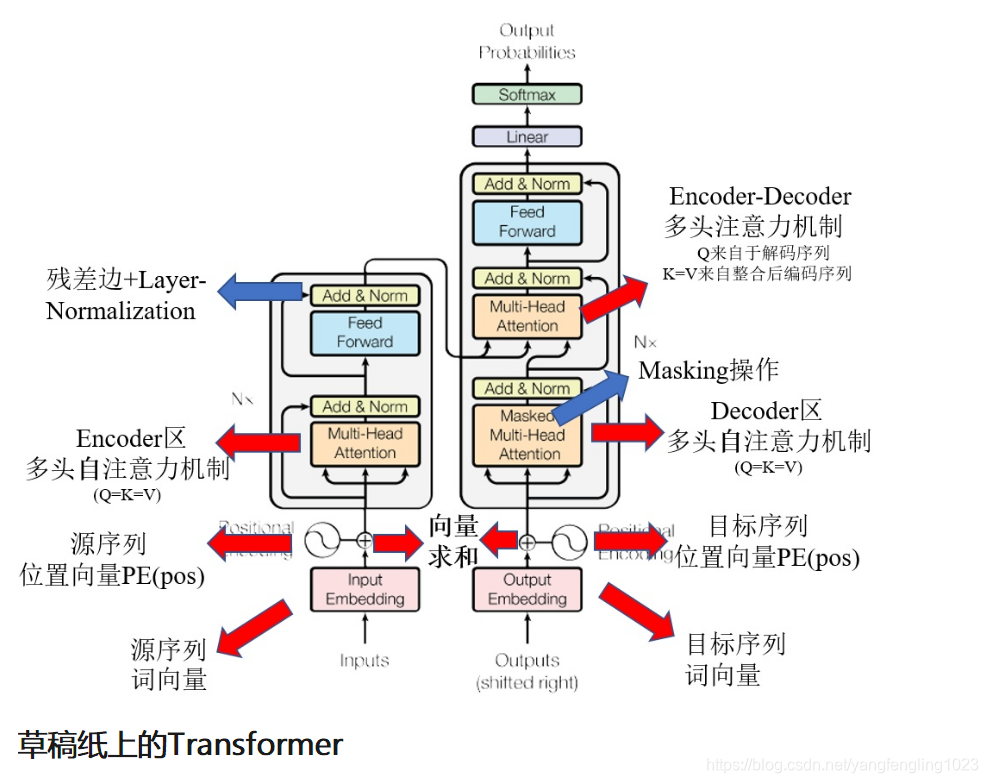
1. 解码器（Decoder）

解码器从编码器的隐藏状态开始，并生成目标序列。与编码器类似，解码器也可以由 RNN、LSTM 或 GRU 构成。每一步解码时，解码器生成一个输出，并将该输出作为下一步的输入。

1. Transformer模型

Transformer 模型是由 Vaswani 等人在 2017 年提出的，最初用于机器翻译任务。与传统的序列到序列模型不同，Transformer 模型完全依赖于注意力机制（Attention Mechanism）而无需循环神经网络（RNN）。这种架构能够更有效地并行处理数据，从而显著加快训练速度，并提高模型在长距离依赖关系上的表现。

Transformer 模型由编码器（Encoder）和解码器（Decoder）两个主要部分组成。每个部分又由多个堆叠的层构成。下图是Transformer的整体架构图。



图二、Transformer模型网络架构

1. 编码器（Encoder）：编码器接收输入序列，并将其转换为隐藏状态表示。编码器由多个相同的层（通常是 6 层）堆叠而成，每层包含几个主要的子层。即Multi-Head Attention、Add & Norm和Feed Forward三个小模块组成。
2. 解码器（Decoder）：解码器从编码器输出的隐藏状态中生成目标序列。解码器也由多个相同的层（通常是 6 层）堆叠而成。其结构与编码器的结构相似，只是在输入端改成了Masked Multi-Head Attention。

**Methodology**

本实验首先对原始数据集进行数据预处理，选取总数据集中的子集作为训练集。然后基于开源的Pytorch库分别编写Seq2Seq模型和Transformer模型。模型训练完毕之后，通过编写的Predict函数，给定开头后生成武侠小说的片段。

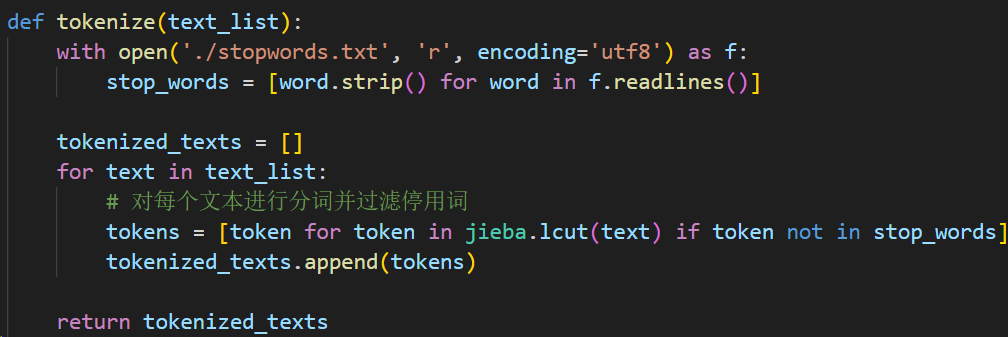
**M1: 数据预处理**

在读取语料后，首先去掉txt文本中一些无意义的广告、无关词语和标点符号等内容。



图三 读取文本数据集函数

此后利用jieba分词对语料进行分词，并读取停用词表进行文本过滤，最后将处理后的训练数据集返回。



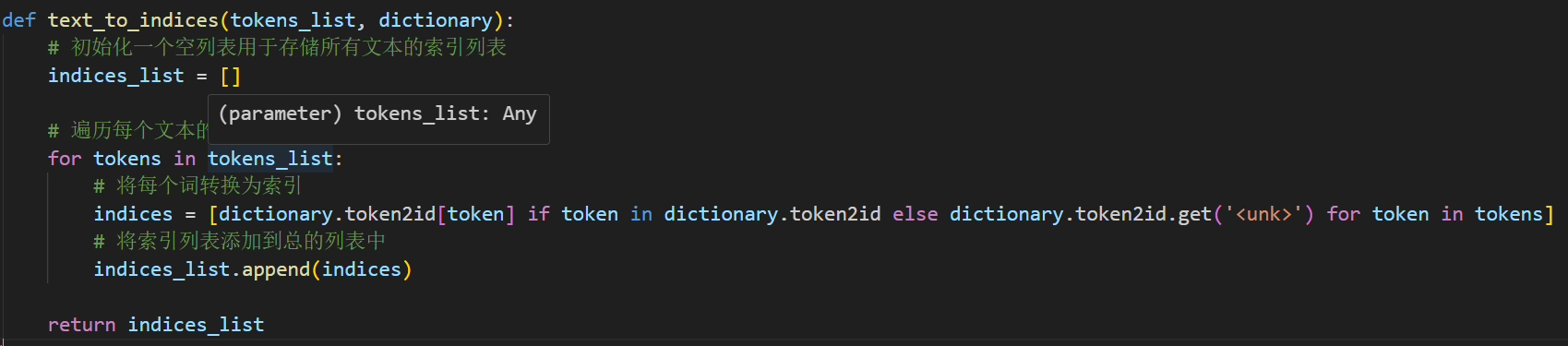
图四 分词函数

此后使用 gensim 库中的 Dictionary 类来创建一个词典实例，统计所有文档中的词汇，并为每个唯一词汇分配一个唯一的ID，并添加了一些特殊标签。此后过滤掉那些在文档中出现次数较少的词汇。



图五 构建词汇表函数

在此之后，将每个单词（token）映射到词典（dictionary）中相应的索引，实现词编码功能。最后基于Pytorch重写供模型进行训练的数据集（Dataset）

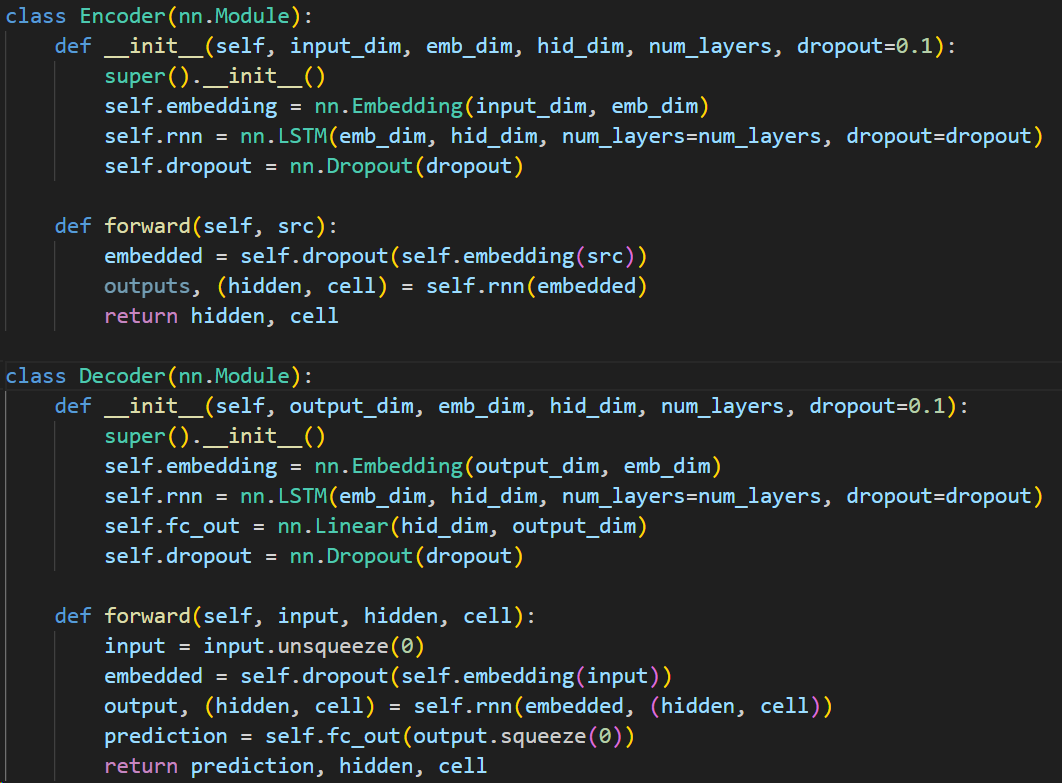


图六 词编码函数

**M2: Seq2Seq模型**

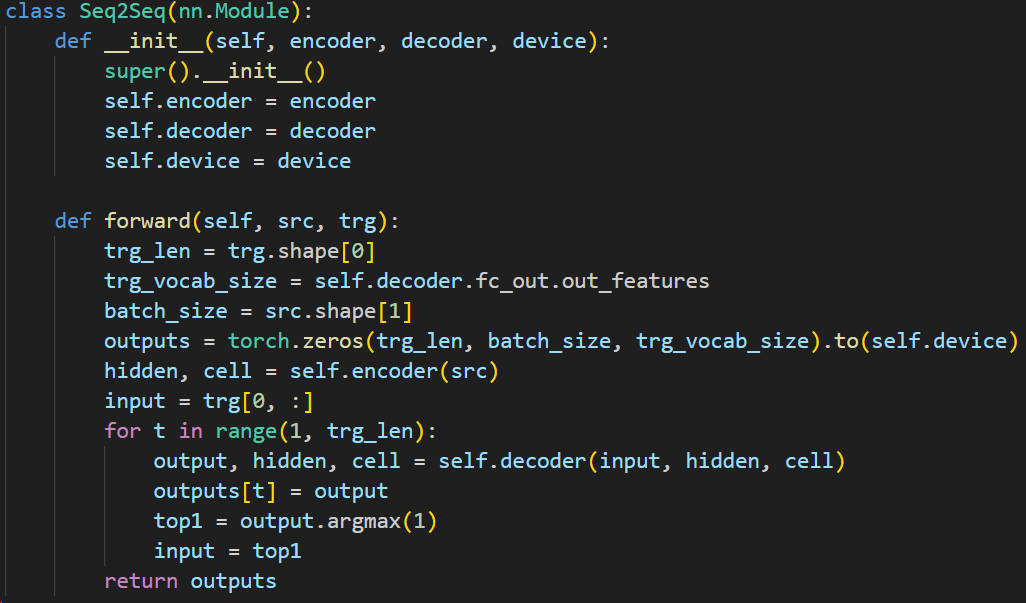
在Introduction部分我们知道Seq2Seq模型由Encoder和Decoder组成。在具体实现上，我们使用LSTM模型完成Encoder和Decoder的构建。

具体来说，Encoder主要由Embedding和LSTM模块组成；Decoder由Embedding、LSTM和线性层（Linear）组成。



图七 Seq2Seq模型中的Encoder和Decoder模块

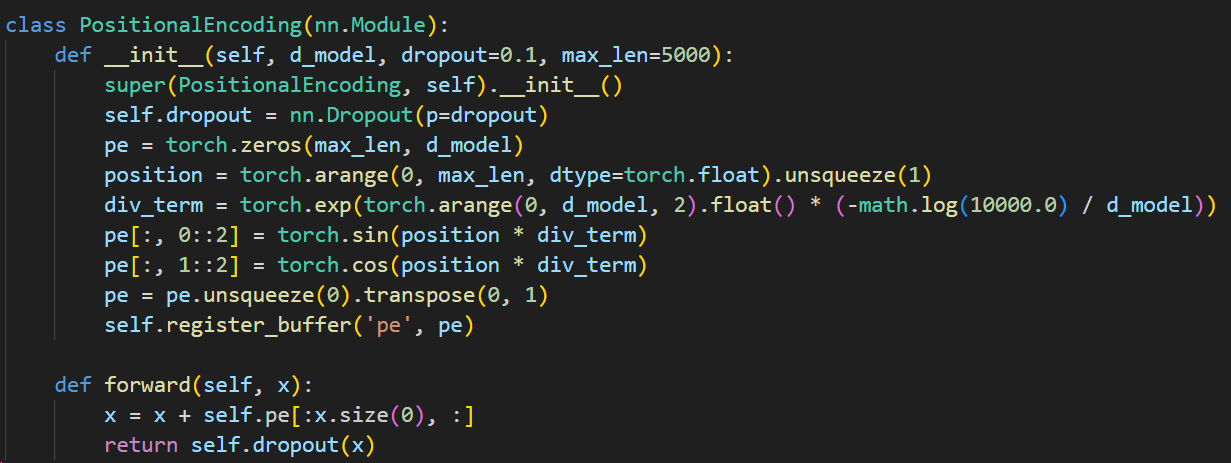
整个Seq2Seq模型使用Teacher Forcing的策略。在使用 Seq2Seq 模型进行训练时，解码器需要逐步生成目标序列的每个时间步。Teacher Forcing 的核心思想是在每个时间步将实际的目标输出（而不是模型的预测）作为解码器的下一步输入。这样可以加速训练收敛，并帮助模型学习更好的目标分布。



图八 Seq2Seq模型框架

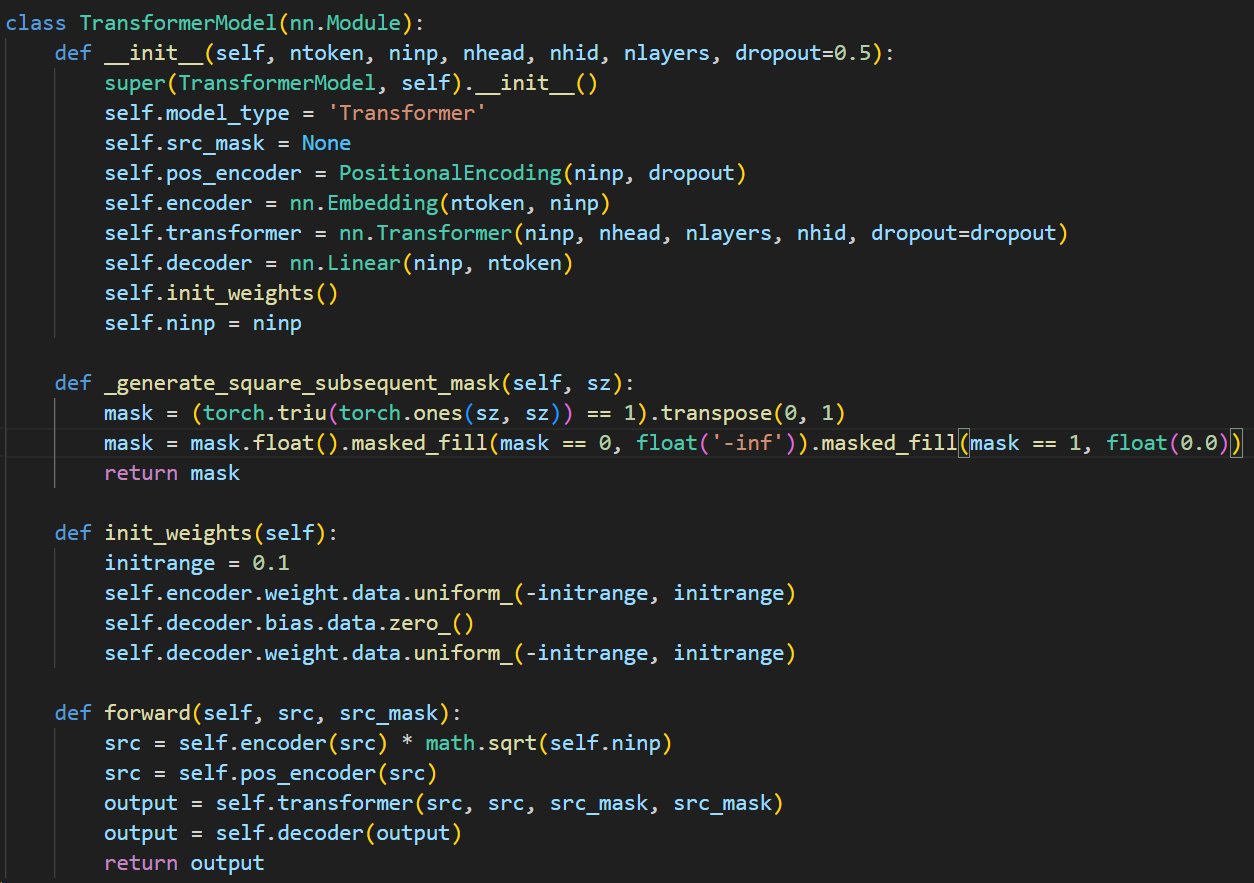
**M3: Transformer模型**

首先实现了Transformer中的位置编码函数。



图九 位置编码函数

然后基于Pytorch中现有的Transformer模块实现Transformer模型，其中包括编码器、解码器、位置编码和多头自注意力等模块。



图十 Transformer模型框架

**Experimental Studies**

由于计算资源和时间有限，本次实验只选取了神雕侠侣作为训练集进行了训练。输入相同的语料后得到的实验结果如下所示。

|  |  |  |
| --- | --- | --- |
| 输入 模型 | Seq2Seq | Transformer |
| 一阵轻柔婉转的歌声，飘在烟水蒙蒙的湖面上。歌声发自一艘小船之中，船里五个少女和歌嘻笑，荡舟采莲。 | 一阵轻柔婉转的歌声，飘在烟水蒙蒙的湖面上。歌声发自一艘小船之中，船里五个少女和歌嘻笑，荡舟采莲。周伯通在过落北阳星，略有数十年雪火，心中不喜全都惊呼的玉箫武功深湛称怪，群雄越来越非，怎能嬴胜让口，将徒母是洞志但江湖衣部尚发石室，此人号令肥胖的道个，必能重阳虽已无约，但对方王旗群道不少年时。 | 一阵轻柔婉转的歌声，飘在烟水蒙蒙的湖面上。歌声发自一艘小船之中，船里五个少女和歌嘻笑，荡舟采莲。节近中秋，荷叶渐残，莲肉饱实。但季节、景物以及越女的容貌、衣著、首饰、心情，下半阕更是写景中有叙事，叙事中挟抒情，自近而远，馀意不尽。 |

通过对比可以发现，由Seq2Seq模型虽然训练时间较短，且生成的语句虽略微带有小说的意味，但是语句较为不通顺；而Transformer模型能够综合上下文信息，推断出非常接近原文的语句，但是训练时间较长，计算资源消耗大。

**Conclusions**

本次实验基于金庸小说语料库分别使用Seq2Seq模型和Transformer模型进行了文本生成的实验。实验结果显示，Seq2Seq模型适用于短文本生成，结构简单清晰，训练速度快，但是生成的语句存在不通顺的现象；而Transformer模型能够综合上下文信息，推断出非常接近原文的语句，但是训练时间较长，计算资源消耗大。

**References**

[1] https://blog.csdn.net/weixin\_42663984/article/details/117068473