Report of Deep Learning for Natural Langauge Processing

Homework_4

Pan Yao
1239388514@qq.com

Abstract

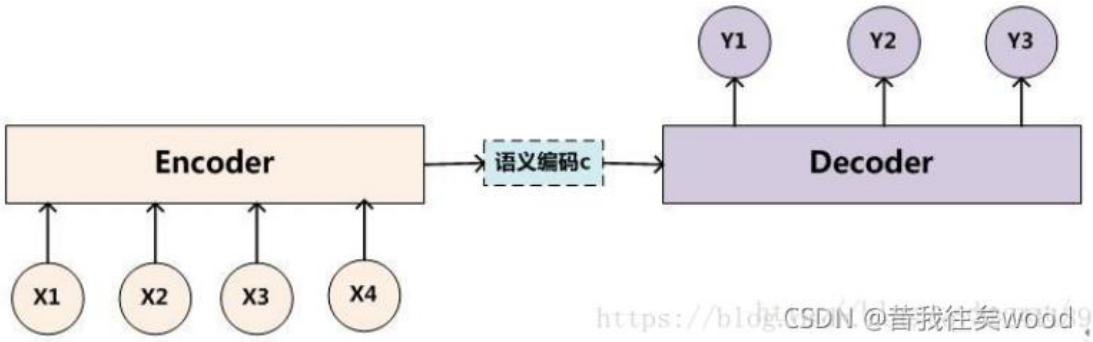
利用给定语料库，用 Seq2Seq 与 Transformer 两种不同的模型来实现文本生成的任务（给定开头后生成武侠小说的片段或者章节），并对比与讨论两种方法的优缺点。

Introduction

一、Seq2Seq

Seq2Seq(Sequence to Sequence)，即序列到序列模型，就是一种能够根据给定的序列，通过特定的生成方法生成另一个序列的方法，同时这两个序列可以不等长。这种结构又叫 Encoder-Decoder 模型，即编码-解码模型，其是 RNN 的一个变种，为了解决 RNN 要求序列等长的问题。

Seq2sSeq 模型其标准结构如下：



Seq2Seq 最基础的结构由三部分组成，编码器，语义向量 C 和解码器，C 是连接二者的。编码器通过学习，将输入序列编码成一个固定大小的状态向量 C，作为解码器的输入，解码器 RNN 通过对 C 的学习进行输出。此处解码器和编码器都代表一个 RNN，通常为 LSTM 和 GRU。

其基本思想利用两个 RNN，一个作为 Encoder，一个作为 Decoder。前者负责将输入的文本序列压缩成指定长度的向量，即语义向量 C，这个向量可以看作输入序列的语义，这个过程成为编码。

编码一般有两种方式，将 RNN 最后一个状态做一个变换得到语义向量，或者将输入序列的所有隐含状态做一个变换得到语义向量。Decoder 负责根据语义向量生成指定的序列，即解决问题的序列，即解码。如下图，最简单的方式是将语义向量 C 作为初始状态输入到 Encoder 的 RNN 中，得到输出序列。此时上一时刻的输出会成为当前时刻的输入，而且语义向量 C 只作为初始状态参与运算，后面运算与 C 无关。第二种方式语义向量 C 参与序列所有时刻的运算。

二、Transformer

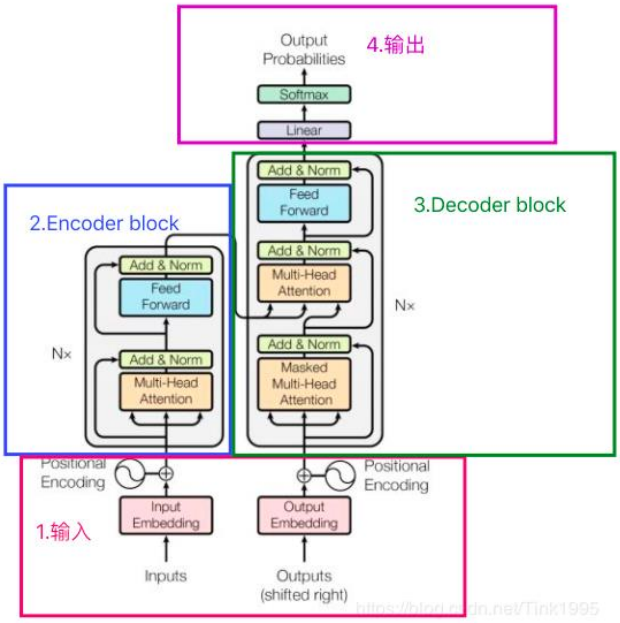
Transformer 模型是由 Google 在 2017 年提出的，旨在解决传统的序列到序列模型在处理长距离依赖问题上的不足。传统的 RNN 和 LSTM 模型在处理长文本序列时，容易出现梯度消失或爆炸问题，导致模型性能下降。Transformer 模型通过引入自注意力机制和多头注意力机制，成功地解决了这一问题。

Transformer 的重要组成部分和特点：

1. 自注意力机制（Self-Attention）：这是 Transformer 的核心概念之一，它使模型能够同时考虑输入序列中的所有位置，而不是像循环神经网络（RNN）或卷积神经网络（CNN）一样逐步处理。自注意力机制允许模型根据输入序列中的不同部分来赋予不同的注意权重，从而更好地捕捉语义关系。
2. 多头注意力（Multi-Head Attention）：Transformer 中的自注意力机制被扩展为多个注意力头，每个头可以学习不同的注意权重，以更好地捕捉不同类型的关系。多头注意力允许模型并行处理不同的信息子空间。
3. 堆叠层（Stacked Layers）：Transformer 通常由多个相同的编码器和解码器层堆叠而成。这些堆叠的层有助于模型学习复杂的特征表示和语义。
4. 位置编码（Positional Encoding）：由于 Transformer 没有内置的序列位置信息，它需要额外的位置编码来表达输入序列中单词的位置顺序。
5. 残差连接和层归一化（Residual Connections and Layer Normalization）：这些技术有助于减轻训练过程中的梯度消失和爆炸问题，使模型更容易训练。
6. 编码器和解码器：Transformer 通常包括一个编码器用于处理输入序列和一个解码器用于生成输出序列，这使其适用于序列到序列的任务，如机器翻译。

Transformer 的结构：

Encoder block 由 6 个 encoder 堆叠而成，图中的一个框代表的是一个 encoder 的内部结构，一个 Encoder 是由 Multi-Head Attention 和全连接神经网络 Feed Forward Network 构成。如下图所示：



Experiment

一、Seq2Seq

基于 LSTM 构建 Seq2Seq 模型：

1. 数据预处理：读取所有文本内容，剔除无用的标点等，并通过 jieba 库将文本进行分词处理。
2. 定义自定义数据集类：用于加载和处理分词后的小说句子。其中 collate_fn：用于在批处理中进行填充，确保所有句子的长度一致。
3. 构建词汇表：统计词频并构建词汇表，包括特殊标记如<pad>、<sos>、<eos>和<unk>。
4. 定义 Encoder：编码器将输入序列嵌入到高维空间，并通过 LSTM 进行处理，输出隐藏状态和细胞状态。
5. 定义 Decoder：解码器接收编码器的隐藏状态和细胞状态，以及前一个时间步的输出，生成下一个时间步的输出。
6. 定义 Seq2Seq 模型：组合编码器和解码器，实现序列到序列的生成。通过 teacher forcing 机制，决定是使用实际目标输入还是使用模型的预测作为下一个时间步的输入。
7. 模型训练：训练模型，使用梯度累积来处理小批量数据。
8. 文本生成：从给定的起始文本生成新的文本。

二、Transformer

采用更适合用于生成中文文本的预训练 GPT-2 模型。

通过 Hugging Face 模型库下“uer/gpt2-chinese-cluecorpussmall”，并下载所需的文件，所需文件如下：

config.json - 包含模型的配置信息，如模型大小、激活函数类型等。

pytorch_model.bin - 包含使用 PyTorch 框架训练的模型权重。

tokenizer_config.json - 包含分词器的配置信息。

vocab.txt - 分词器使用的词汇表。

special_tokens_map.json - 包含特殊标记的映射，例如开始标记、结束标记等。

Experimental Results

文本生成：

在 epoch=10，max_length=300 的条件下，实验结果如下：

给定的初始文本：青衣剑士回剑侧身，右腿微蹲，锦衫剑士看出破绽，挺剑向他左肩疾刺。不料青衣剑士这一蹲乃是诱招，长剑突然圈转，直取敌人咽喉，势道劲急无伦。锦衫剑士大骇之下，长剑脱手，向敌人心窝激射过去。这是无可奈何同归于尽的打法，敌人若是继续进击，心窝必定中剑。当此情形，对方自须收剑挡格，自己便可摆脱这无可挽救的绝境。不料青衣剑士竟不挡架闪避，手腕抖动，噗的一声，剑尖刺入了锦衫剑士的咽喉。

学习率 lr=5e-5 时，Seq2Seq 的生成文本：女子 从未 浙江 也 不敢 一刹那 一刹那 印出来 印出来 得 得 诵 诵 诵 诵 诵 诵 诵 实在 印出来 印出来 印出来 印出来 攻击 生动 印在 溪 其中 印在 溪 竹名 竹名 简陋 竹梢 竹梢 竹梢 浙江 之野 极少数 极少数 共有 人 同时 简陋 向来 向来 向来 是 美丽 不敢 再 再 作 挑 问以 反正 第二篇 不敢 第二篇 不敢 如 再 现代 现代 吾 实在 实在 实在 印出来 印出来 给 给 给 没有 没有 没有 很 处女 很 生死 生死 生死 生死 生死 生死 生死 生死 其中 生死 生死 其中 溪 其中 印在 溪 竹名 竹名 简陋 竹梢 竹梢 竹梢 浙江 之野 极少数 极少数 共有 人 同时 简陋 向来 向来 向来 是 美丽 不敢 再 再 作 挑 问以 反正 第二篇 不敢 第二篇 不敢 如 再 现代 现代 吾 实在 实在 实在 印出来 印出来 给 给 给 没有 没有 没有 很 处女 很 生死 生死 生死 生死 生死 生死 生死 生死 其中 生死 生死 其中 溪 其中 印在 溪 竹名 竹名 简陋 竹梢 竹梢 竹梢 浙江 之野 极少数 极少数 共有 人 同时 简陋 向来 向来 向来 是 美丽 不敢 再 再 作 挑 问以 反正 第二篇 不敢 第二篇 不敢 如 再 现代 现代 吾 实在 实在 实在 印出来 印出来 给 给 给 没有 没有 没有 很 处女 很 生死 生死 生死 生死 生死 生死 生死 生死 其中 生死 生死 其中 溪 其中 印在 溪 竹名 竹名 简陋 竹梢 竹梢 竹梢 浙江 之野 极少数 极少数 共有 人 同时 简陋 向来 向来 向来 是 美丽 不敢 再 再 作 挑 问以 反正 第二篇 不敢 第二篇 不敢

如 再 现代 现代 吾 实在 实在 实在 印出来 印出来 给 给 给 没有 没有 没有 很 处女 很 生死 生死 生死 生死 生死 生死 生死 生死 其中 生死 生死 其中 溪 其中 印在 溪 竹名 竹名 简陋 竹梢 竹梢 竹梢 浙江 之野 极少数 极少数

Transformer 的生成文本:第 一 回 合 双 方 僵 持 了 一 会 儿 后 ， 锦 衫 剑 士 开 大 门 进 了 他 的 房 间 ， 对 方 见 他 衣 衫 神 色 凝 重 ， 怕 他 进 去 不 便 脱 身 而 进 。 他 心 中 嘀 咕 不 已 ， 不 由 得 长 叹 一 声 ， 只 见 锦 衫 剑 士 左 胸 口 猛 地 一 扬 ， 锦 衫 剑 士 竟 伸 缩 右 腿 ， 直 取 敌 人 咽 喉 ， 当 时 已 是 深 夜 ， 这 时 长 剑 已 脱

由实验结果可知，在此实验中，Transformer 模型表现出色，生成的文本更连贯且有逻辑性,而 Seq2Seq 模型生成的文本存在重复和无意义的内容。这表明，尽管 Seq2Seq 模型结构简单且易于实现，但在处理复杂的自然语言生成任务时，Transformer 模型具有明显的优势。

Seq2Seq 的优点：

(1) 语法结构：生成的文本中一些部分保持了句子的结构，尽管重复较多。

缺点：

(1) 内容重复：生成的文本中存在大量的重复词汇和短语，显示出模型在生成过程中出现了循环。

(2) 缺乏上下文连贯性：生成的文本缺乏逻辑和上下文的连贯性，难以理解和应用。

(3) 长距离依赖性差：模型在捕捉长距离依赖关系时表现不佳，导致生成的内容不连贯。

Transformer 的优点：

(1) 高质量生成文本：从生成的文本可以看出，Transformer 生成的内容更连贯，有逻辑性，避免了明显的重复和无意义的内容。

(2) 并行计算：自注意力机制允许并行计算，大大提高了训练和推理的效率。

(3) 处理长距离依赖：Transformer 模型能够有效地捕捉输入序列中的长距离依赖关系，生成更连贯的文本。

缺点：

(1) 模型复杂：Transformer 模型结构复杂，包括多头注意力机制、位置编码等，需要更深的理解和更多的实现细节。

(2) 计算资源要求高：Transformer 模型通常需要大量的计算资源，对硬件要求较高。

由于硬件条件和时间限制，Seq2Seq 模型和 Transformer 模型均未能取得理想的效果，其中 Seq2Seq 模型的生成效果较差，后续需要进行参数的调整以及代码的优化。