Report of Deep Learning for Natural Langauge Processing4

Yongquan Chen
cyqss@buaa.edu.cn

Abstract

在这个任务中，本文利用金庸小说语料库，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），针对这两种模型，本报告分别使用 Seq2Seq 模型以及基于 Transformer 的 GPT-2 模型实现文本生成，并对比与讨论两种方法的优缺点。

Introduction

Seq2Seq 模型通常由两个主要部分组成：编码器（Encoder）和解码器（Decoder）。编码器处理输入序列，将其编码为固定长度的上下文向量（context vector）。解码器根据上下文向量生成输出序列。它的工作流程是输入序列通过编码器，每个时间步（时间步 t）都会更新隐藏状态；编码器的最后一个隐藏状态作为上下文向量，传递给解码器。解码器根据上下文向量和之前生成的词（在训练时为真实词，在推理时为上一步生成的词），逐步生成输出序列。

Transformer 模型由编码器堆栈（Encoder Stack）和解码器堆栈（Decoder Stack）组成，每个堆栈包含多个相同的层（Layers）。每个编码器层包含自注意力机制（Self-Attention）和前馈神经网络（Feed-Forward Neural Network）。每个解码器层包含自注意力机制、编码器-解码器注意力机制（Encoder-Decoder Attention）和前馈神经网络。

Transformer 模型执行流程是输入序列通过嵌入层（Embedding Layer）并添加位置编码（Positional Encoding）以保留顺序信息。然后编码器堆栈对输入序列进行编码，生成一系列隐藏状态。最后解码器堆栈根据编码器的输出和前一步的输出序列（在训练时为真实词，在推理时为上一步生成的词），逐步生成输出序列。

Methodology

对给定的金庸数据集，本文首先将这些文本合并为一个大文本文件。接着，使用 Jieba 库对文本进行分词，并根据句末标点进行分句，以获得清晰明确的分句分词语料。

随后分别加载 Seq2Seq 模型和 GPT-2 模型并对模型进行训练，由于后者训练时间过长本文也没有多少计算资源，因此本文直接下载预训练完成的模型和参数，并在完成训练的模型上再使用相关语料库进行训练，虽然减少了实际效果，但是效果相对可看。

Experimental Studies

对两个模型提供相同的示例输入，观察他们的输出情况。

Generated text by Seq2Seq：少年下山游历，七八个七八个七八个七八个七八个七八个七八个七八个七八个七八个七八个七八个七八个七八个七八个七八个七八个

Generated text by Transformer (GPT-2): 少年下山游历，来一场说走就走的旅行武侠小说中，有一称得上是侠之大者的人物，他们的出场，定是为了寻找一个真正的侠客。他们的出场，定是为了寻找一个真正的侠客。他们的出

Conclusions

很容易看出两者的效果都不太好，相对而言 GPT-2 效果更好些。

比较两个模型：

Seq2Seq 模型：适合处理中小规模的任务，结构简单，计算效率相对较低，但在处理序列和捕捉长程依赖关系方面存在一定的局限性。

Transformer 模型：适合大规模任务，具有更高的计算效率和更好的长程依赖捕捉能力，但需要更多的计算资源和更高的技术理解能力。观察到现在的 GPT-4o 的效果，可以看出后者的发展潜力更大。