作业报告四

Weijia Long
wjial@buaa.edu.cn

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），针对这两种模型，本报告分别使用基于 Seq2Seq 的 T5 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

Part 1

首先处理中文语料库，并先根据 jieba 进行中文分词，在根据不同模型的 tokenizer 进行提取信息，如代码中 def preprocess_chinese_corpus(file_path):与 def load_dataset_from_corpus(corpus, tokenizer):所示，通过数据预处理得到训练集。

随后加载模型和分词器并对模型进行训练，由于模型训练时间过长且需要计算资源，因此本报告采用微调的方式使用中文语料库对下载好的模型进行微调，如代码微调函数 def finetune_model(model, train_dataset, output_dir, tokenizer)所示。

最后使用模型生成文本并观察效果如代码 def generate_text(model, tokenizer, prompt, max_length=50, num_beams=3)所示。

```
# 示例生成文本
prompt = "武林至尊"
print("Generated text by Seq2Seq (T5):", generate_text(t5_model, t5_tokenizer, prompt))
print("Generated text by Transformer (GPT-2):", generate_text(gpt2_model, gpt2_tokenizer, prompt))
```

Experimental 1

在训练完模型后，经过分析结果可得出 Seq2Seq（T5）具有以下优点：

一、序列到序列架构： Seq2Seq 模型使用编码器-解码器结构，适用于处理输入和输出序列长度不同的任务，如机器翻译、摘要生成等。二、可解释性： Seq2Seq 模型生成的文本是通过解码器逐步生成的，每一步都可以理解为一个基于当前上下文的决策，因此具有一定的可解释性。

但其缺点也很明显，首先是生成质量不稳定， T5 模型可能需要进行调参和后处理以获得更好的效果。（实现的实验结果并不是很好）。

GPT-2 的优点就是语言理解能力较好， GPT-2 模型生成的文本不仅具有良好的语法结构，还具有一定的语义理解能力，可以根据上下文生成合理的文本回复，具有更广泛的适用性和更强的语言理解能力。

实验结果：

Generated text by Seq2Seq (T5): 师父个个无无无无无无无无无无无无无无无无无无无无无无无无无无无无无无无无无无

Generated text by Transformer (GPT-2): 武林至尊 青年 关明梅女 那时 崇 玄冥 玄冥女 还是 局 方不到 名 哑 对羊 地 这个 此处 陡然 心道 猴儿镖局 叔 摸 投 干老 镖局叔 摸 听见 兵 每 干 丘

至少看起来 GPT-2 更能生成有用的信息

T5 训练过程：

{'train_runtime': 13.5325, 'train_samples_per_second': 0.222, 'train_steps_per_second': 0.222, 'train_loss': 9.959003448486328, 'epoch': 3.0}

GPT-2 训练过程：

{'train_runtime': 6.1178, 'train_samples_per_second': 0.49, 'train_steps_per_second': 0.49, 'train_loss': 10.033636728922525, 'epoch': 3.0}