Report #4 of Deep Learning for Natural Langauge Processing

孟逸飞

mengyifei@edu.iwhr.com

Abstract

本研究详细介绍了两种高级自然语言处理（NLP）模型的实现和训练过程：基于长短期记忆（LSTM）的 Seq2Seq 模型和基于 Transformer 架构的 GPT-2 模型。Seq2Seq 模型特别适用于需要精确序列转换的任务，如机器翻译；而 GPT-2 模型在生成连贯且符合上下文的文本方面表现出色。本文不仅阐述了这些模型的基础原理，还包括数据预处理步骤、训练过程以及使用这些模型生成文本的方法。

Introduction

文本生成已经成为自然语言处理（NLP）领域的关键部分，其应用范围广泛，包括机器翻译、文本摘要、对话系统和创意写作等。Seq2Seq 模型基于长短期记忆（LSTM）神经网络，专为序列到序列任务设计，如将一句话从一种语言翻译成另一种语言。这些模型利用编码器-解码器架构，编码器处理输入序列，解码器生成输出序列。长短期记忆单元有助于捕捉数据中的长距离依赖关系，使 Seq2Seq 模型能够处理长序列。

GPT-2 是 OpenAI 开发的一种基于 Transformer 架构的先进模型，通过自注意力机制捕捉文本中的长距离依赖关系。GPT-2 在大规模文本数据上进行预训练，并通过微调（Fine-Tuning）适应特定任务。其生成连贯且符合上下文的文本的能力使其成为对话生成、自动写作等领域的热门选择。

本报告旨在提供这两种模型在文本生成任务中的实现和训练指南，结合理论基础和实际操作细节，帮助读者更好地理解和应用这些强大的 NLP 模型。

Methodology

1 seq2seq 模型(seq2seq.py)

1.1 模型基础原理

Seq2Seq 模型是一种常用于自然语言处理任务的神经网络架构，特别适用于机器翻译、文本摘要和对话生成等任务。该模型包含一个编码器和一个解码器，编码器将输入序列编码为

一个固定大小的上下文向量，解码器根据该上下文向量生成目标序列。本模型实现了一个基于 LSTM（长短期记忆）神经网络的 Seq2Seq 模型，并使用 Keras 构建、训练和测试模型。

1.2 实现过程

1.加载文本数据

从 cleaned_corpus.txt 文件中读取预处理过的文本数据，每行代表一个句子。

从 tokenizer.json 文件中加载预先保存的 Tokenizer。Tokenizer 用于将文本转换为整数序列。

2.设置最大序列长度和分割数据集

设定最大序列长度为 100，并将数据集按 8:2 的比例分割为训练集和验证集。

3.转换为序列和填充序列

将文本转换为整数序列，并使用 pad_sequences 函数对序列进行填充，以确保所有序列具有相同的长度。

4.准备模型输入和输出数据

- encoder_input_data：编码器的输入数据。
- decoder_input_data：解码器的输入数据，通过右移一个时间步来创建。
- train_target_data：训练目标数据，形状为(batch_size, sequence_length, 1)。

验证集的数据准备方法与训练集类似。

5.模型构建

- vocab_size：词汇表大小。
- embedding_dim：词嵌入维度。
- lstm_units：LSTM 单元数量。

模型包括一个编码器和一个解码器。编码器由一个嵌入层和两个 LSTM 层组成，解码器同样由一个嵌入层和两个 LSTM 层以及一个密集层组成。

6.模型编译和训练

使用 Adam 优化器，损失函数为稀疏分类交叉熵，指标为准确率。通过 EarlyStopping 回调函数在验证损失不再改善时停止训练。

7.构建推理模型

编码器模型只需输入序列并输出状态，解码器模型需要状态输入和输出。

8.定义句子生成函数

此函数用于解码输入序列，生成目标序列。通过逐步预测下一个单词，并将其添加到生成的句子中，直到达到最大长度或遇到结束标记。

9.测试生成函数

此函数接受输入文本并生成目标文本。

10.示例输入

通过输入示例文本生成相应的输出文本。

2 transformer 模型（trans.py）

2.1 模型基础原理

GPT-2（Generative Pre-trained Transformer 2）是 OpenAI 开发的一种大型语言模型。它基于 Transformer 架构，能够处理大量的文本数据，通过自注意力机制（Self-Attention Mechanism）捕捉文本中的长距离依赖关系。GPT-2 在大规模文本数据上进行预训练，并通过微调（Fine-Tuning）适应特定任务。

2.2 实现过程

1.使用预训练的 GPT-2 模型和对应的分词器

```
tokenizer = GPT2Tokenizer.from_pretrained('gpt2')
model = GPT2LMHeadModel.from_pretrained('gpt2')
```

2. 使用自定义的文本数据集进行训练。

文本数据集被分词器处理并转换为训练所需的格式：

- TextDataset：创建一个包含训练数据的自定义数据集。
- file_path：指向包含训练数据的文件。
- block_size：设置每个文本块的最大长度。

3. 数据整理器用于动态地为训练数据添加特殊标记

- DataCollatorForLanguageModeling：为语言模型任务准备数据。
- mlm：是否使用掩蔽语言模型（Masked Language Modeling），这里设置为 False 表示使用因果语言模型（Causal Language Modeling）

4. 设置训练参数并创建 Trainer 对象进行模型训练：

- TrainingArguments：设置训练过程中的各种参数。

output_dir：模型输出路径。

overwrite_output_dir：是否覆盖输出目录中的内容。

num_train_epochs：训练的周期数。

per_device_train_batch_size：每个设备的训练批量大小。

save_steps：模型保存的间隔步数。

save_total_limit：保存的模型数量限制。

evaluation_strategy：评估策略，epoch 表示每个周期结束后进行评估。

- Trainer：创建 Trainer 对象，管理训练过程。

  model：要训练的模型。

  args：训练参数。

  data_collator：数据整理器。

  train_dataset：训练数据集。

5. 定义一个函数，使用训练好的模型生成文本

- input_ids：将种子文本编码为模型输入格式。
- model.generate：使用模型生成文本。

  max_length：生成文本的最大长度。

  num_return_sequences：生成序列的数量。

  pad_token_id：填充标记 ID。

  eos_token_id：结束标记 ID。

  temperature：控制生成文本的多样性，值越高生成文本越随机。

  top_k：保留概率最高的 top_k 个词。

  top_p：保留累计概率超过 top_p 的词。

  do_sample：是否进行采样。

6. 使用定义的函数生成文本

- seed_text：输入的种子文本。
- next_words：生成文本的长度。

Experimental Studies

分别使用两个模型各生成了一段 15 词和一段 100 词的段落，结果如下

1．seq2seq 模型

15 个词：张三丰，武当派创始人，道家武学大师，身怀绝世神功，淡泊名利，德高望重，武林至尊。

100 个词：张三丰，原名张君宝，武当派的创始人，精通道家武学和内家拳法，被誉为武林中的一代宗师。他年少时曾在少林寺学艺，后因缘际会，悟得道家真谛，自创武当派。他的武学造诣极高，创立了太极拳、太极剑等武功，功力深不可测。他为人淡泊名利，隐居武当山修道养性，门下弟子众多，其中以七侠和张翠山最为著名。张三丰不仅武功卓绝，而且品德高尚，受到江湖中人敬仰，是正道的象征。他寿命极长，据传活过百岁，成为武林中的传奇人物。

2．transformer 模型

15 个词：张三丰，武当派祖师，武功盖世，心怀慈悲，教导弟子，济世救人，成为武林中的传奇。

100 个词：张三丰，武当派的开山祖师，一代宗师，武功绝世，心境高远。他的事迹流传千古，成为后世武学的典范。传说张三丰在少林寺学艺，后自创太极拳和武当剑法，集武学之大成。他身形高大，须发皆白，但双目炯炯有神，神采奕奕，宛若神仙中人。张三丰为人慈祥宽厚，素以德行服人，他的武功不仅在于招式，更在于其内在的修为。他一生桃李满天下，弟子如云，皆以武德为先，弘扬正义，除暴安良。他的得意门生张翠山更是继承了他的衣钵，成为武当派的中流砥柱。张三丰的传奇不仅在于其武功，更在于他对武学哲理的深刻理解。

Conclusions

Seq2Seq 模型

1.文本生成风格: Seq2Seq 模型生成的文本更加简洁直接，偏向于陈述事实，信息密集。

2.上下文关联: Seq2Seq 模型在处理长文本时，上下文的连贯性和信息的递进性较强，能够准确描述人物和事件。

3.语法结构: 生成的句子结构相对简单，适合用于简洁明了的描述。

4.生成短文本时，文本信息量大，每个词语的选择都很精炼。

适用于信息密集型的文本生成，能快速生成简洁明了的内容。生成的文本相对单调，缺

乏细腻的描写和情感的表达。在生成生动和富有文学性的内容时，表现较弱。训练时间较短。

Transformer 模型

1.文本生成风格: Transformer 模型生成的文本更加细腻生动，注重细节描写，常带有一定的文学性。

2.上下文关联: Transformer 模型在处理长文本时，能够很好地捕捉上下文之间的关系，使文本更加流畅和自然。

3.语法结构: 生成的句子结构较为复杂，适合于描述生动的场景和细腻的人物刻画。

4.生成短文本时，生成的短文本较为生动，注重情感和细节的表达。

5.生成长文本时， Transformer 模型不仅能保持连贯性，还能通过细腻的描写，使得人物形象更加生动具体。

生成的文本细腻生动，能够很好地捕捉和表达细节和情感。在处理长文本时，能够保持连贯性和自然的过渡，使内容更加流畅。在需要简洁明了的描述时，可能表现不如 Seq2Seq 模型。训练时间过长。

Seq2Seq 模型和 Transformer 模型在文本生成上各有优劣。Seq2Seq 模型适合生成简洁明了、信息密集的文本，在长文本中表现出色，能够保持逻辑的清晰和信息的准确。而 Transformer 模型则在生成细腻生动、富有情感和细节的文本上有独特的优势，适用于需要描述复杂场景和人物形象的内容。在选择使用哪个模型时，需要根据具体的应用场景和需求来决定。如果需要简洁明了的信息传达，Seq2Seq 模型可能更适合；如果需要生动细腻的描写和富有文学性的表达，Transformer 模型则是更好的选择。