

Transformer 翻译实验报告

2023200440

2024 年 12 月 5 日

1 实验目的

本次实验通过补充 Transformer 翻译的实现，加深对 Transformer 的理解，初步了解机器学习中的翻译任务。并且能够对实现过程的一些细节尝试解释。

2 实验环境

机器：windows11

解释器：python3.12

编辑器：pycharm

3 实验过程

3.1 Task1: 实现 self-attention 的计算

首先计算分数矩阵。query, key 和 value 是对同一个输入序列的线性变换表示，三者维度相同。首先计算 query 维度 d_k ，然后计算 query 和 key 的相似性分数，这里需要将 key 的最后两个维度转置，从而得到序列中每个元素与其他所有元素的相似度分数矩阵，并除以维度 d_k ，进行缩放，避免 softmax 中梯度爆炸或者消失的情况。

计算 query 和 key 的维度

```
d_k = query.size(-1)
```

将 query 和 key 的最后一个维度进行转置后相乘，再除以缩放因子

```
scores = torch.matmul(query, key.transpose(-2, -1)) / math.sqrt(d_k)
```

然后根据 mask 来实现注意力掩码机制, 将 mask 为 0 的位置将分数值设置为负无穷, 这些位置将不会对注意力权重产生贡献。

```
if mask is not None:
    scores = scores.masked_fill(mask == 0, float('-inf'))
```

计算注意力权重, 并对 v 进行加权 (已给出)

```
# 使用 softmax 函数计算注意力权重
p_attn = F.softmax(scores, dim=-1)
```

3.2 Task2: 实现 Transformer 的 forward

首先将源语言序列 src 和源语言序列掩码 src_mask 输入编码器得到编码表示, 即解码器的 memory。然后将 memory, 源语言序列的掩码 src_mask, 目标语言序列的掩码 tgt_mask 输入解码器, 对 tgt 进行嵌入, 然后进行解码, 返回输出。

```
def forward(self, src, tgt, src_mask, tgt_mask):
    # encoder 的结果作为 decoder 的 memory 参数传入, 进行 decode
    # 编码器的输出
    memory = self.encode(src, src_mask)
    # 解码器的输出
    output = self.decode(memory, src_mask, tgt, tgt_mask)
    # 返回解码器的输出
    return output
```

3.3 Task3: 输入英文分词和 word2id

首先将输入的英文句子使用 word_tokenize 进行分词, 并加上开始标记"BOS" 和结束标记"EOS"。然后根据词典 data.en_word_dict, 获取单词索引, 单词不存在则返回 UNK, 最后得到单词索引列表。

```
sentence_en = ["BOS"] + word_tokenize(sentence_en) + ["EOS"]
sentence_en_ids = [data.en_word_dict.get(word, UNK) for word in sentence_en]
```

3.4 question 1

Question:

为什么 `trg` 舍弃最后一个字符, `trg_y` 舍弃第一个字符?

Answer:

`trg` 是解码器的输入, 训练时我们不希望解码器看到"EOS", 因为解码器应该预测下一个词, 而"EOS" 是已经给出的结束标记, 不需要预测"EOS"的下一个, 因为"EOS" 意味着当前序列已经结束了。`trg_y` 是解码器的目标输出, 它不包括序列的开始标记"BOS", 因为解码器的第一个输出就是针对"BOS" 的预测。

3.5 question 2

Question:

训练过程和测试过程 decoder 的输入有何不同?

Answer:

训练过程: - decoder 的输入是目标序列的真实值, 包括"BOS" 和"EOS", 每个时间步输入是真实值或者"BOS" 和"EOS"。

测试过程: - decoder 的输入是前一时间步的预测值或者是开始符号(BOS)。例如, 如果第一个时间步的预测是'我', 那么在第二个时间步, decoder 的输入是'我', 直到预测出 EOS 或者达到最大长度。

4 实验结果

4.1 结果展示

实验结果如下:

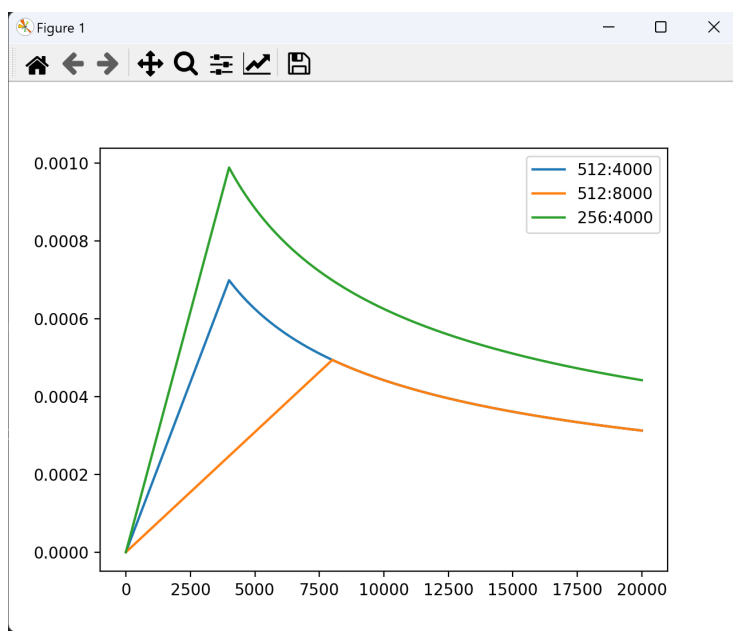


图 1: 学习率曲线

```
src_vocab 5493
tgt_vocab 2537
>>>>>> Training starts!

Epoch 0 Batch: 113 Loss: 5.6314: 100% ██████████ 114/114 [01:35<00:00, 1.19it/s]
>>>> Evaluating on dev set...
Epoch 0 Batch: 14 Loss: 6.0674: 100% ██████████ 15/15 [00:14<00:00, 1.03it/s]
<<<<< Evaluate loss: 5.831693
***** Save model done! *****

Epoch 1 Batch: 113 Loss: 4.5873: 100% ██████████ 114/114 [01:28<00:00, 1.28it/s]
>>>> Evaluating on dev set...
Epoch 1 Batch: 14 Loss: 5.1623: 100% ██████████ 15/15 [00:10<00:00, 1.41it/s]
<<<<< Evaluate loss: 4.926178
OK | 0/114 [00:00<?, ?it/s]***** Save model done! *****

Epoch 2 Batch: 113 Loss: 3.9228: 100% ██████████ 114/114 [01:29<00:00, 1.28it/s]
>>>> Evaluating on dev set...
Epoch 2 Batch: 14 Loss: 4.6156: 100% ██████████ 15/15 [00:10<00:00, 1.38it/s]
<<<<< Evaluate loss: 4.323038
***** Save model done! *****

Epoch 3 Batch: 113 Loss: 3.3618: 100% ██████████ 114/114 [01:40<00:00, 1.14it/s]
>>>> Evaluating on dev set...
Epoch 3 Batch: 14 Loss: 4.0112: 100% ██████████ 15/15 [00:11<00:00, 1.26it/s]
<<<<< Evaluate loss: 3.744939
OK | 0/114 [00:00<?, ?it/s]***** Save model done! *****

Epoch 4 Batch: 113 Loss: 2.8862: 100% ██████████ 114/114 [01:32<00:00, 1.23it/s]
>>>> Evaluating on dev set...
Epoch 4 Batch: 14 Loss: 3.5932: 100% ██████████ 15/15 [00:08<00:00, 1.74it/s]
<<<<< Evaluate loss: 3.299782
***** Save model done! *****

<<<<<<< Training finished. Cost 524.9572 seconds.
D:\CODE_REPOSITORY\AI\0\0\0\0\2023280640_9\0\lab71\transformer-nmt-pub\transformer-nmt.py:319: FutureWarning: You are using 'torch.load
```

图 2: 训练结果

```

请输入英文: The boy was naked to the waist.
译文: 这个男孩子们的车。

请输入英文: She doesn't speak Japanese at home.
译文: 她不是个人不是她的人。

请输入英文: I'm staying at the Hilton Hotel.
译文: 我在这个男孩子。

请输入英文: He earns his living by writing.
译文: 他是他的人学生。

请输入英文:

```

图 3: 翻译示例

5 实验分析

从翻译示例来看，训练得到的模型进行翻译的效果并不好，翻译不准确，可能的原因分析如下：

5.1 数据集大小

训练集数据只有一万五左右条中英语句，相比较商用的翻译功能，这个数据量是非常小的，保存的词典和捕获的语义语法是比较有限的。因此翻译效果明显不好。并且翻译时也可以发现，很多词语在词典中并不存在。

```

请输入英文: Springing to my feet, I received my first Martian surprise.
BOS UNK to my feet , I received my first UNK surprise . EOS
译文: 我的时候我的时候，他的我很多时候我的。

请输入英文: For greater sharpness, but with a slight increase in graininess
BOS for greater UNK , but with a slight increase in UNK EOS
译文: 这个孩子在一个人有一个人都是一个人都有一个人都有一个人都是一个人都是一个人都是一个人。

```

图 4: missword

5.2 模型参数

除此以外，与模型超参数的设置也有关系。实验初始设置的训练轮数是 5，但是可以看到训练结果中损失并没有降到很低（大致在 3 到 5 之间），说明模型拟合效果不是很好，实际翻译效果也确实如此。数据集和注意力头的

个数也有关系。如果数据集不够大，过多的注意力头可能会导致模型过拟合，因为模型可能会学习到数据中的噪声。

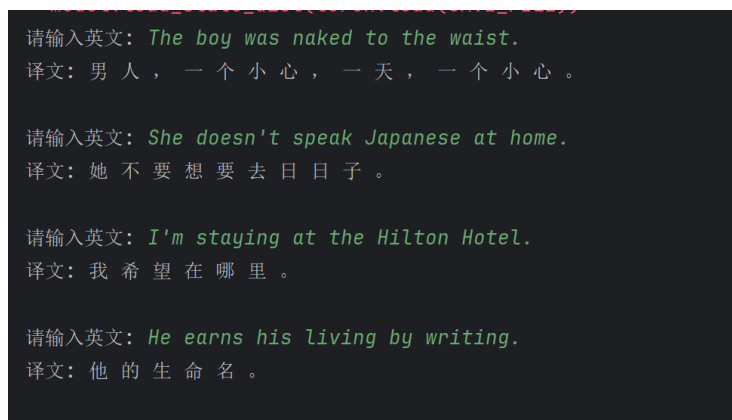
5.3 客观原因

乔姆斯基提出的乔姆斯基层次结构也说明了“由有限状态机表示的语法成为有限状态语法但它不能够生成所有句子”，当输入一些比较新的语义和语法结构时，基于原本语料的翻译模型，对于预料中没有出现的句法结构翻译效果可能不是那么好。

6 实验改进

6.1 增加数据集大小

我从飞桨平台下载了一个约 520 万个中英文平行语料集，截取了 5 万条用于训练，训练后的效果如下：



```
请输入英文: The boy was naked to the waist.  
译文: 男人, 一个小心, 一天, 一个小心。  
  
请输入英文: She doesn't speak Japanese at home.  
译文: 她不要想去日子。  
  
请输入英文: I'm staying at the Hilton Hotel.  
译文: 我希望在哪里。  
  
请输入英文: He earns his living by writing.  
译文: 他的生命名。
```

图 5: 增大数据集翻译示例

从结果来看，效果没有明显变化，这与训练轮数设置有关。在 5 万数据集下，设置 5 轮训练花费了一个小时左右。不过从神经网络的经验来看，增大数据集是一个改善模型性能的有效途径。

6.2 增加训练次数

基于原来的模型，我将训练轮数增加到了 20, 发现损失明显降低，并且翻译效果显著优化。

```
Epoch 17 Batch: 113 Loss: 0.9189: 100% | 114/114 [01:08<00:00, 1.65it/s]
>>>> Evaluating on dev set...
Epoch 17 Batch: 14 Loss: 0.6086: 100% | 15/15 [00:07<00:00, 1.89it/s]
<<<< Evaluate loss: 0.747726
***** Save model done! *****

Epoch 18 Batch: 113 Loss: 0.9002: 100% | 114/114 [01:08<00:00, 1.66it/s]
>>>> Evaluating on dev set...
Epoch 18 Batch: 14 Loss: 0.5276: 100% | 15/15 [00:07<00:00, 1.89it/s]
<<<< Evaluate loss: 0.649792
***** Save model done! *****

Epoch 19 Batch: 113 Loss: 0.7702: 100% | 114/114 [01:09<00:00, 1.65it/s]
>>>> Evaluating on dev set...
Epoch 19 Batch: 14 Loss: 0.4388: 100% | 15/15 [00:07<00:00, 1.90it/s]
<<<< Evaluate loss: 0.560636
***** Save model done! *****

<<<<<< Training finished. Cost 1543.5543 seconds.
```

图 6: 增加训练轮数损失

```
请输入英文: He knows better than to marry her.
译文: 他 比 她 更 愿 意 和 她 结 婚 。

请输入英文: He had hoped to succeed, but he didn't.
译文: 他 曾 经 希 望 能 取 得 他 并 不 得 更 努 力 。

请输入英文: This is the worst movie I have ever seen.
译文: 这 是 我 曾 看 过 的 电 影 。

请输入英文: She's in the bath.
译文: 她 在 浴 缸 里 唱 个 歌 。

请输入英文: |
```

图 7: 增加训练轮数集翻译示例

7 实验总结

总的来说，这是一个十分值得探索的实验。不过，由于时间原因不能充分探究。目前来看，对于模型参数调整和更大的数据集训练，以及对模型的更深入学习从而优化模型结构是几个值得探索的方向。实验也缺少合理的定性和定量评估标准来测试模型性能和效果，这仍待改进。实验中也尝试了增

加数据集大小和训练轮数的方法尝试改进模型效果，这对加强 Transformer 模型的理解和在下游任务中的运用是很有益的。