组会汇报

陈轲杰
专业:计算数学

June 21, 2023

目录

对instruction类型的key-value进行格式修改

```
def format_example(example: dict) -> dict:
context = f"Instruction: {example['instruction
    ']}\n"
if example.get("input"):
    context += f"Input: {example['input']}\n"
context += "Answer: "
target = example["output"]
# {"context": context, "target": target}
example['context'] = context
example['target'] = target
return example
```

- 将instruction和input进行合并作为context
- 将output作为target

数据集的形式

- 对于基本的时间序列中形如”2021-01-01 00:00:00”,在词长度等上面比较繁琐，所以使用一个简单的序列x来替代
- 使用了如下的基本序列：$\begin{bmatrix} x & sin(x) \\ 1 & sin(1) \\ \vdots & \vdots \end{bmatrix}$
- 指令微调的形式如下：

```
{
    "Instruction": "For a test data sequence, from the existing 16 sequences,
        predict the last column of future 8 sequences?",
    "Input": "
        0,0.0,#,1,0.8415,#,2,0.9093,#,3,0.1411,#,4,-0.7568,#,5,-0.9589,#,6,-0.2794,
        ",
    "Output": "-0.2879,-0.9614,-0.751,0.1499,0.9129,0.8367,-0.0089,-0.8462"
}
```

- 其中预测形式是(input:1-16 output:17-20),(input:2-17 output:18-22),...

对测试集合的调整

- 在数量上一共给定了10000和30000长度的序列。
- 在instruction预测长度有三种：16预测8,8预测4,两种混合的情况(这种情况下序列长度是20000)
- 两种序列：单变量序列$(x, sin(x))$,多变量序列$(x, cos(x), sin(x^2+2), sin(x))$
- 使用的预训练模型为chatglm

记10000长度，16预测8,单变量序列的数据集的训练结果：result-10000-16-8-one

Rouge评价指标

- 举个例子说明召回率，精确度，F1分数三种指标，举个例子说明：比如有200件信封，其中垃圾邮件150，现在要判定好坏邮件，模型判断出了100封有问题的邮件。但是实际上有80封是垃圾邮件，其他的是没有问题的。
  1. 召回率就是正确预测的数量/所有正确的数量=80/150=0.53，
  2. 精确度就是正确预测的数量/所有预测为正的样本数量=80/100=0.8
  3. F1分数是综合考虑召回率和精确度的评估指标。他是召回率和精确度的调和平均值，用于评估模型在分类任务中的性能。
     计算公式：$\text{F1分数} = \frac{2(\text{精确度} * \text{召回率})}{\text{精确度} + \text{召回率}} = 0.637$
  4. F1 分数的取值范围是0 到1，越接近1 表示模型的性能越好。

评价指标

- Rouge-1（Rouge-N）：衡量生成的摘要与参考摘要之间的unigram（单个词）重叠程度。
- Rouge-2（Rouge-N）：衡量生成的摘要与参考摘要之间的bigram（两个连续词）重叠程度。
- Rouge-L：衡量生成的摘要与参考摘要之间的长序列重叠程度。
- BLEU-4：计算4个连续词之间的重叠程度
- (rouge指标通常使用上述的F1分数)

运行结果展示

- 见result.xlsx

运行结果分析

- 基本上预测长度越短，效果就越好，比如根据8个预测后面4个效果比根据16个数据预测后面8个数据要更好。
- 使用的序列越长效果反而不好，猜测可能的原因是发生了过拟合。
- 数据集中包含不同预测长度的序列的模型的预测结果良好。

后续工作

后续工作

1. 前面测试的虽然是不同的预测长度序列，但是数据集的序列是固定的，比如是(x,sin(x))序列不同的长度的预测。之后可以尝试同一个数据集里有(x,sin(x)),(x,cos(x),sin(x))等多种不同的序列，并且预测长度也可以不同。
2. 改变评估指标
   这边的评估指标Rouge是文本摘要质量的评估指标,这个指标主要评估的是序列之间的相似度，比如目标序列1.1,1.2,1.3,预测结果100,1.2,1.3的效果是远高于预测结果1.0,1.3,1.4。所以对于序列预测，最终的评估指标应该改成MAE或MSE.构建合适的评估数据集。
3. 当前的目标：此模型能够对于已有的多种不同的时间序列数据集(已经见过的)进行预测。

谢谢老师和同学的聆听!