NLP hw3

杨舒

1 Task1

1.1 LLM 推理加速

在这个任务中，我对比了 gpt2 默认设置、使用 kv-cache、使用不同程度量化（不使用 kv-cache）的吞吐率和内存分配。

模型输入为"Discuss the ethical implications of gene editing technologies like CRISPR, focusing on its potential impact on biodiversity."，模型输出最长长度为 100.

我的硬件配置是 NVIDIA GeForce GPU，拥有 24GB 显存，运行 CUDA 12.0，驱动版本为 525.60.13。

结果如下：

Type	**Throughput (tokens/sec)**	**Allocated Memory (bytes)**	**Time (s)**
Baseline	64.45	520,323,072	1.55
KV-cache	176.62	520,323,072	0.57
float16 Quantized	176.49	280,742,912	0.57
int8 Quantized	77.66	702,216,704	1.29
int4 Quantized	126.69	675,458,048	0.79

Table 1: Performance metrics for different configurations

其中 kv-cache 可以增加推理速度，但不能减少内存占用，使用.half() 方法进行 float6 量化可以加快推理速度并减少内存消耗，值得注意的是，使用 bitsandbytes 进行 int8 和 int4 量化虽然可以增加推理速度，但是反而会增加内存分配，可能和使用的硬件有关。

1.2 Custom KV-Cache

Batch Size	**Time for Golden(s)**	**Time for Customized(s)**	**Improvement Ratio**
16	90.35	65.12	1.42
32	88.43	62.19	1.42
48	88.75	61.93	1.43
64	94.48	61.50	1.54

Table 2: Comparison of Time Taken for Golden Greedy and Customized Greedy Decoding with Different Batch Sizes

自定义 KV-Cache 的实现可以增加推理速度，并且增大 batchsize 后提升更明显。

2 LLM 推理技巧

2.1 设置

在这个任务中，我使用 GSM8K 数据集测试集的 1319 条数据进行准确率的测试。使用的模型为 deepseek-chat，模型其他设置均为默认 (temperature = 1.0)。

我的 Naive Prompt 为：Answer the following question: (question)；我的 CoT Prompt 为：Answer the following question: (question) Let's think step by step；我的 In-context Learning Prompt 为：Given the following examples:(context)Now, answer this question: (question)，其中 context 的内容为 gsm8k 训练集的前三个问题-答案配对；我的 few-shot CoT Prompt 为：Given the following examples:(context)Now, answer this question: (question) Let's think step by step.

在模型通过 prompt 输出结果后，还会再接受一个指令："So your calculated number is:"

Reflexion 的具体流程是：使用和上文中 CoT 相同的 prompt 生成答案，与标准答案对比验证答案，如果正确则完成；如果错误则将失败的结果添加到反思历史中，并且生成新的反思内容，并将其添加到反思字符串中，供下一轮使用。使用 CoT 和历史所有的反思内容生成新答案，一共可以尝试 3 次。

2.2 准确率比较

Table 3: Model Accuracy for Different Prompts

Prompt Type	Accuracy
Naive Prompt	0.9075
CoT Prompt	0.9121
ICL Prompt	0.9242
Few-shot CoT Prompt	0.9257
Reflexion	0.9600

2.3 分析

不同 prompt 方法相对于 Naive Prompt 都可以对准确率有所提升，其中 Reflexion 的提升最大。In-context Learning 总体来说效果比 CoT 好，但从具体例子来看这两个方法相比 Naive Prompt 增加的正确的例子不太相同。Few-shot CoT 相对 In-context Learning 的提升很小。

以下是具体的例子：

Problem 1: Naive, COT Unsolved; ICL Solved

Question: Judy teaches 5 dance classes every day on weekdays and 8 classes on Saturday. If each class has 15 students and she charges $15.00 per student, how much money does she make in 1 week?

Answer: She teaches 5 dance classes 5 days a week, so that’ s:

$$5 \times 5 = 25 \,\text{classes}$$

She teaches 25 classes during the week and 8 classes on Saturday, for a total of:

$$25 + 8 = 33 \,\text{classes}$$

There are 15 students in each of the 33 classes, so the total number of students is:

$$15 \times 33 = 495 \,\text{students}$$

Each student pays $15.00 per class, so Judy makes:

$$15 \times 495 = 7,425$$

icl 给出的例子的思维链长度与这一题类似，比较适合这一题，而它的思维链长度可能对于简单 CoT 来说较长，使得模型不能很好跟进每一步的结果。

Problem 2: Naive, ICL Unsolved; CoT Solved

Question: Brandon’s iPhone is four times as old as Ben’s iPhone. Ben’s iPhone is two times older than Suzy’s iPhone. If Suzy’ s iPhone is 1 year old, how old is Brandon’ s iPhone?

Answer: Ben’ s iPhone is:

$$1 \times 2 = 2 \,\text{years old}$$

Brandon’ s iPhone is:

$$4 \times 2 = 8 \,\text{years old}$$

这个例子推导步骤很短，可能是因为 icl 给出的例子步骤都相对更长，所以没能提高这一题的效果。但较短的思维链条很适合简单的 CoT Prompt。

Problem 3: Naive, COT, ICL Unsolved; Few-Shot Solved

Question: Jill gets paid $20 per hour to teach and $30 to be a cheerleading coach. If she works 50 weeks a year, 35 hours a week as a teacher and 15 hours a week as a coach, what’s her annual salary?

Answer: First, find the total amount Jill makes per week teaching:

$$20 \times 35 = 700 \,\text{USD/week}$$

Then find the total amount Jill makes per week coaching:

$$30 \times 15 = 450 \,\text{USD/week}$$

Then add these two amounts to find the total amount Jill makes per week:

$$700 + 450 = 1,150 \,\text{USD/week}$$

Multiply this by the number of weeks Jill works in a year to find her annual salary:

$$1,150 \times 50 = 57,500 \,\text{USD}$$

icl 和 CoT 结合在一起，一定程度可以弥补这两者的不足，使模型能够解决推理步骤稍多、计算种类稍多的问题。

Problem 4: Only Reflexion Solved

Question: Josh decides to try flipping a house. He buys a house for $80,000 and then puts in $50,000 in repairs. This increased the value of the house by 150%. How much profit did he make?

Answer: The cost of the house and repairs came out to:

$$80,000 + 50,000 = 130,000\,\text{USD}$$

He increased the value of the house by:

$$80,000 \times 1.5 = 120,000\,\text{USD}$$

So the new value of the house is:

$$120,000 + 80,000 = 200,000\,\text{USD}$$

Thus, he made a profit of:

$$200,000 - 130,000 = 70,000\,\text{USD}$$

这个问题中涉及的数字比较大，可能不适合模型的推理任务，而 Reflexion 通过让模型多次作答并使用反思信息的方式使模型突破了原本能力的限制。