LLMs 推理性能面

来自： AiGC面试宝典

宁静致远

2024年01月27日 19:44

一、介绍一下 LLMs 的文本生成过程?

LLMs 的文本生成过程 分为：

1. **【预填充（prefill）】**阶段：以并行方式处理输入提示中的词元；
2. **【解码（decoding）】阶段：文本会以自回归的方式逐个生成“词元”。**每个生成的词元都会被添加到输入中，并被重新喂入模型，以生成下一个词元。**当LLM输出了特殊的停止词元或满足用户定义的条件（例如生成了最大数量的词元）时，生成过程就会停止**

二、如何准确衡量模型的推理速度呢?

1. **首个词元生成时间（Time To First Token，简称TTFT）**：即用户输入查询后，模型生成第一个输出所需的时间。在实时交互中，低时延获取响应非常重要，但在离线工作负载中则不太重要。此指标受处理提示信息并生成首个输出词元所需的时间所驱动；
2. **单个输出词元的生成时间（Time Per Output Token，简称TPOT）**：为每个查询系统的用户生成一个输出词元所需的时间。这一指标与每个用户对模型“速度”的感知相关。例如，TPOT为100毫秒/词元表示每个用户每秒可处理10个词元，或每分钟处理约450个词，这一速度远超普通人的阅读速度；
3. **时延**：模型为用户生成完整响应所需的总时间。整体响应时延可使用前两个指标计算得出：时延 = （TTFT）+ （TPOT）*（待生成的词元数）；
4. **吞吐量**：推理服务器在所有用户和请求中每秒可生成的输出词元数。

这些指标 如何评估好坏了，可以记住这句话：

- 目标：**以最短的时间生成首个词元、达到最高吞吐量以及在最短的时间内生成输出词元。**

模型能够尽可能快地为尽可能多的用户生成文本

注：需要权衡吞吐量和每个输出词元的时间：与依次运行查询相比，如果我们同时处理16个用户查询，吞吐量会更高，但会花费更长的时间为每个用户生成输出词元。

三、如果对整体推理时延有具体目标，有哪些有效的启发式方法来评估模型?

- **输出长度决定了整体响应时延**：对于平均时延，通常只需将预期/最大的输出词元长度与模型的每个输出词元的整体平均时间相乘；
- **输入长度对性能来说影响不大，但对硬件要求至关重要**：在MPT模型中，添加512个输入词元增加的时延要少于生成8个额外输出词元的时延。然而，支持长输入的需求可能使模型难以部署。例如，建议使用A100-80GB（或更新版本）来为最大上下文长度为2048个词元来部署MPT-7B模型；

- **整体时延与模型大小呈次线性关系**：在相同的硬件上，较大的模型速度较慢，但速度比不一定与参数数量比相匹配。MPT-30B的时延约为MPT-7B时延的2.5倍，LLaMA2-70B的时延约为LLaMA2-13B时延的2倍。