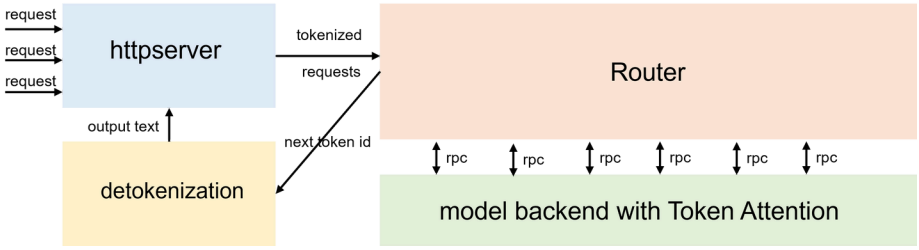
A Light and Fast Inference Service for LLM



LightLLM：轻量高速的LLM

黄浴
自动驾驶话题下的优秀答主

关注他

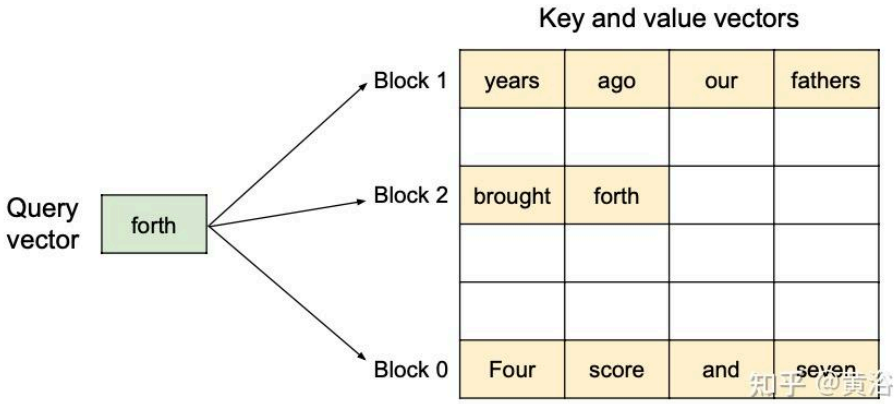
49 人赞同了该文章

LightLLM是一个基于Python的LLM（大型语言模型）推理和服务框架，以其轻量级设计、易于扩展和高速性能而闻名：

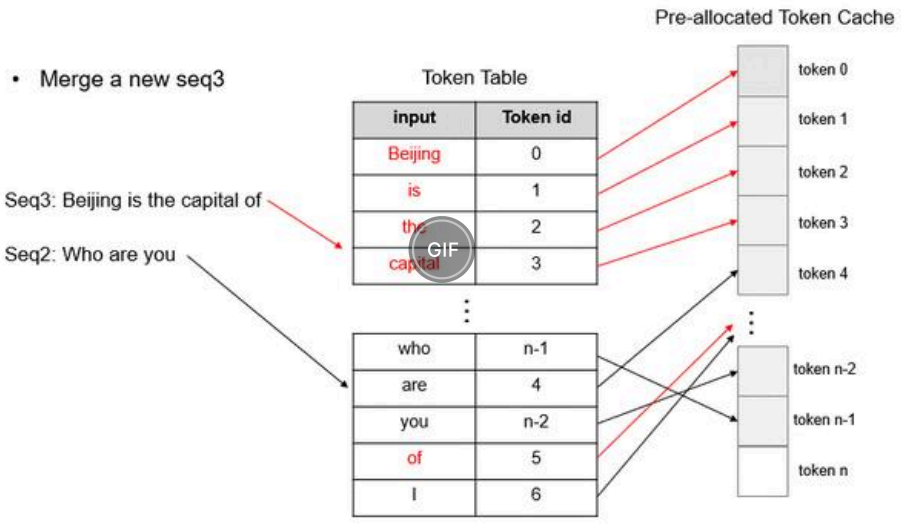
https://github.com/ModelTC/lightllm
github.com/ModelTC/lightllm

LightLLM利用了许多备受好评的开源实现优势，包括FasterTransformer、TGI、vLLM和FlashAttention等。

vLLM中采用的**PagedAttention**将KV缓存存储在不连续的内存空间中。虽然PagedAttention在一定程度上缓解了内存碎片，但仍然为内存浪费留出了空间。此外，在处理多个高并发请求时，内存块的分配和释放效率低下，导致内存利用率不理想。



TokenAttention，一种在token级别管理Key和Value缓存的注意机制。与PagedAttetion相比，TokenAttention不仅最大限度地减少了内存碎片，实现了高效的内存共享，而且有助于高效的内存分配和释放。它允许更精确和细粒度的内存管理，从而优化内存利用率。



下表是PagedAttetion和TokenAttention二者的比较：

Features	PagedAttention	TokenAttention
Low memory fragmentation	✓	✓
Efficient memory sharing	✓	✓
Efficient memory allocation and deallocation	×	✓
Fine-grained memory management	×	✓

知乎@黄浴

由于自注意的时间和内存复杂性在序列长度上是二次型的，因此Transformer在长序列上是缓慢的，并且需要更多内存。近似注意方法试图通过权衡模型质量来降低计算复杂度来解决这个问题，但通常无法实现挂钟（wall-clock）加速。一个缺失的原理是使注意算法具有IO意识——考虑GPU内存级别之间的读写。**FlashAttention**是一种IO-觉察的精确注意算法，用平铺（tiling）来减少GPU高带宽存储器⁺（HBM）和GPU片上SRAM之间的存储器读/写次数。将FlashAttention扩展到块稀疏注意，可产生一种比任何现有近似注意方法都快的近似注意算法。

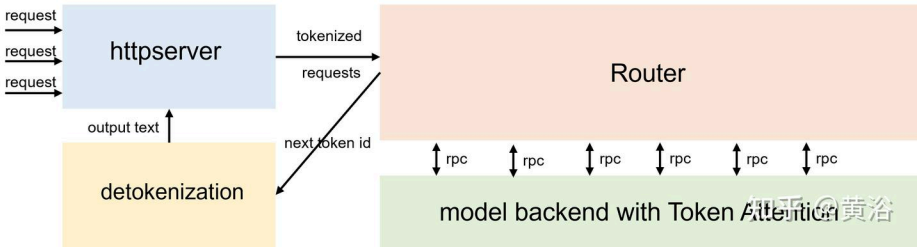
LightLLM包括以下特点：

1. 三进程异步协作：token化⁺、模型推理和去token化是异步执行的，从而大大提高了GPU的利用率。
2. Nopad（Unpad）：支持跨多个模型的无填零注意操作，有效处理长度差异较大的请求。
3. 动态批处理⁺：请求的动态批处理调度。
4. **FlashAttention**：结合FlashAttention（“**FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness “**，2022**)** ，提高速度并减少推理过程中GPU内存占用。
5. 张量并行性⁺：利用多个GPU的张量并行性进行更快的推理。
6. **TokenAttention**：实现逐Token的KV缓存内存管理机制，允许在推理过程中零内存浪费。
7. 高性能路由器⁺：与TokenAttention合作，精心管理每个token的GPU内存，从而优化系统吞吐量。
8. Int8KV缓存：此功能将使token的容量几乎增加一倍。只有LLAMA支持。

如图所示是Light LLM的基本框架：

LIGHT LLM

A Light and Fast Inference Service for LLM



如下表是vLLM和lightLLM的速度比较：

vLLM	LightLLM
Total time: 361.79 s Throughput: 5.53 requests/s	Total time: 188.85 s Throughput: 10.59 requests/s

知乎 @黄治

编辑于 2024-05-23 08:23 · IP 属地美国

内容所属专栏

大模型的技术和应用
语言大模型，视觉-语言模型，多模态大模型，世界模型
订阅专栏

「真诚赞赏，手留余香」

赞赏

还没有人赞赏，快来当第一个赞赏的人吧！

轻量级 LLM 高速缓存

理性发言，友善互动

2 条评论 默认 最新

佛系架构师东方锡
LightLLM相比TensorRT LLM有哪些特点呢?
2024-03-23 · 江苏
回复 喜欢

jiewen
请教一下不同粒度的内存管理，可以按照所以token(batch)，部分token(block)，单token分，对算子读取数据性能影响有多少?
2023-10-09 · 四川
回复 喜欢

推荐阅读