HW10

Tguo67

Question(s):
Read the related materials, and answer the following questions:
1. LServe proposes a unified sparse attention framework. How does it work for the prefill and decoding phase, respectively?
2. What are the two sparse patterns that substantially contribute to the attention score, according to SampleAttention? What are their semantics?
3. How is CRA defined in SampleAttention? Why is it an indicator of the accuracy? How is it approximately determined?

Related Materials:
LServe paper: https://arxiv.org/pdf/2502.14866Links to an external site.
SampleAttention paper: https://arxiv.org/pdf/2406.15486Links to an external site.

Answers:

1. LServe proposes a unified sparse attention framework. How does it work for the prefill and decoding phase, respectively?

- **Unified block-sparse kernel**: both stages use the same tile/block formulation—compute a Tq * Tk tile and skip whole KV blocks judged as not important and this reduces the number of sequential KV iterations rather than micro-sparsifying within an iteration. In prefill Tq larger than 1 in decoding there's one query so Tq is still 1.
- **Prefill (static sparsity + fused kernel)**: LServe offline partitions heads about 50% are turned into streaming heads with a fixed pattern (attend to sinks + local window). Dense and streaming heads are fused in one sparse kernel, KV for streaming vs. dense heads is kept in separate paged caches.
- **Decoding (dynamic page sparsity + static heads)**: Keep the static streaming heads, and on dense heads select only a constant number of KV pages per query via a hierarchical, query-centric selector; run attention against the shortened page table. This bound decoding complexity and combines multiplicatively with quantized KV.

2. What are the two sparse patterns that substantially contribute to the attention score, according to SampleAttention? What are their semantics?

- **Local window pattern:** high scores concentrate near each token—captures recent/short-range context (recency).

- **Column-stripe pattern:** a few columns (keys) attract high attention across many queries—captures global/key contextual anchors ("salient tokens" shared across rows).

3. How is CRA defined in SampleAttention? Why is it an indicator of the accuracy? How is it approximately determined?

- **Definition**: Cumulative Residual Attention (CRA) is the minimum (over queries) of the sum of remaining attention probabilities after sparsification.
- **Why an accuracy indicator**: theory shows near-lossless sparse attention has a CRA lower bound. And empirically, accuracy remains near full-attention when CRA is kept high, so CRA serves as the sparsity quality target.
- **Approximate determination (runtime, low-overhead):**
  This two-stage, query-guided K/V filtering includes
  - stride sample a few queries to get sampled attention rows
  - column-wise reduce those samples to pick top-k key columns per head that satisfy the CRA threshold;
  - merge with a fixed local window mask for compute.
  - Hyperparameters (CRA threshold, sampling ratio, window ratio) are set via light offline profiling.