cuTensor-HT: High Performance Third-order Hierarchical Tucker Tensor Decomposition on GPUs

Hao Huang ,Tao Zhang , Xiao-Yang Liu

School of Computer Engineering and Science, Shanghai University, Shanghai, China
Department of Electrical Engineering, Columbia University, USA
Shanghai Institute for Advanced Communication and Data Science, Shanghai, China

Motivations

- Real data are often modeled as low-rank tensors.
- Tensor operations are compute-intensive.
- The running time and complexity grow rapidly with tensor order and size.

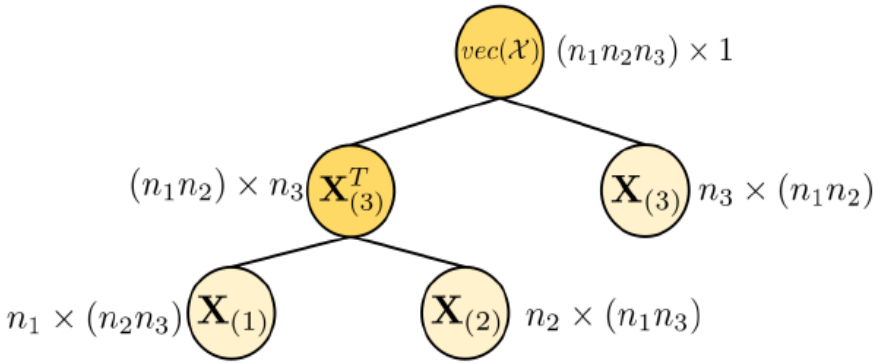
Contributions

We implement the HT decomposition on the GPU.

We propose optimization strategies for memory access,reducing the amount of calculation and improving resource utilization

We conduct experiment to evaluate the performance of HT decomposition algorithm on the GPU and we achieved 4.67 times speedups over the unoptimized GPU baseline.

Parallel HT Decomposition on the GPU



Algorithm 2 Optimized hierarchical Tucker tensor decomposition on GPUs

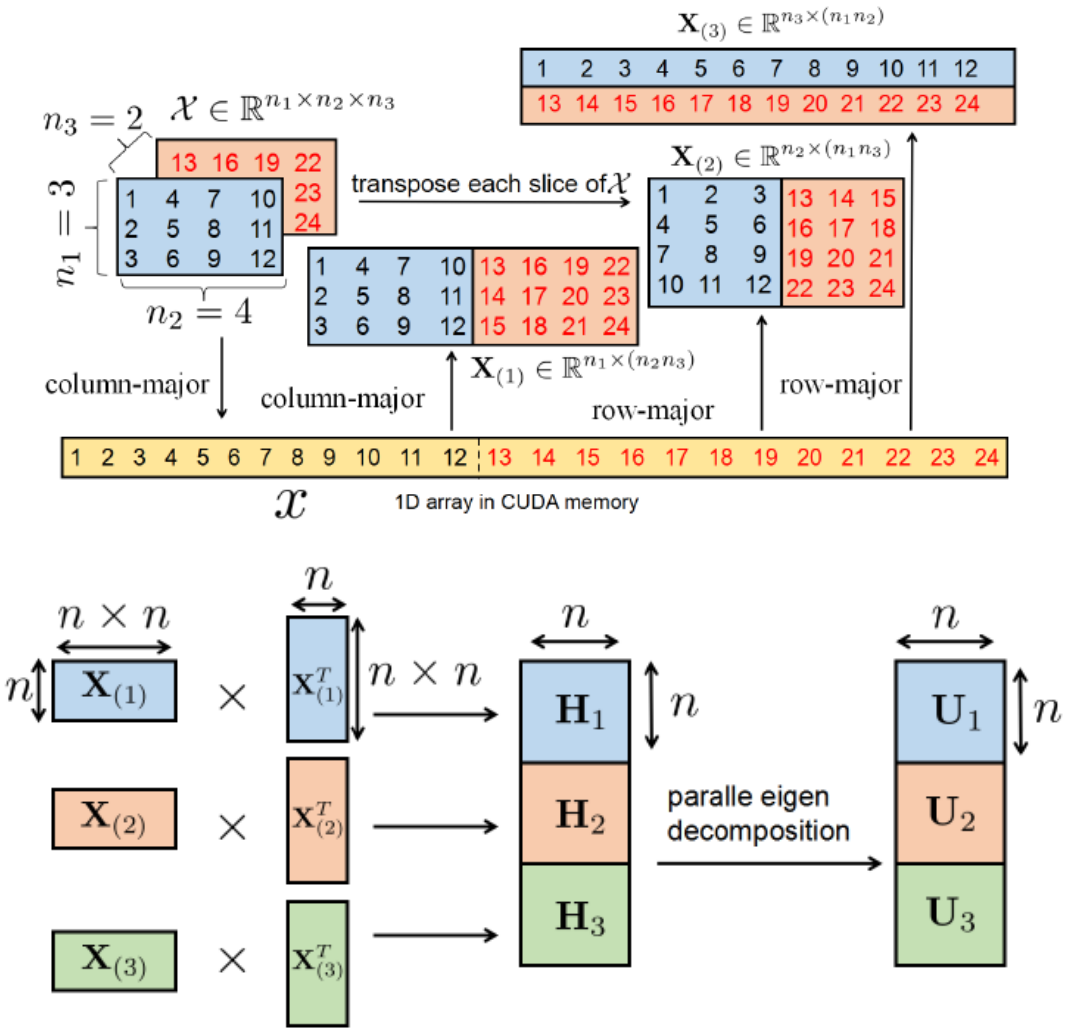
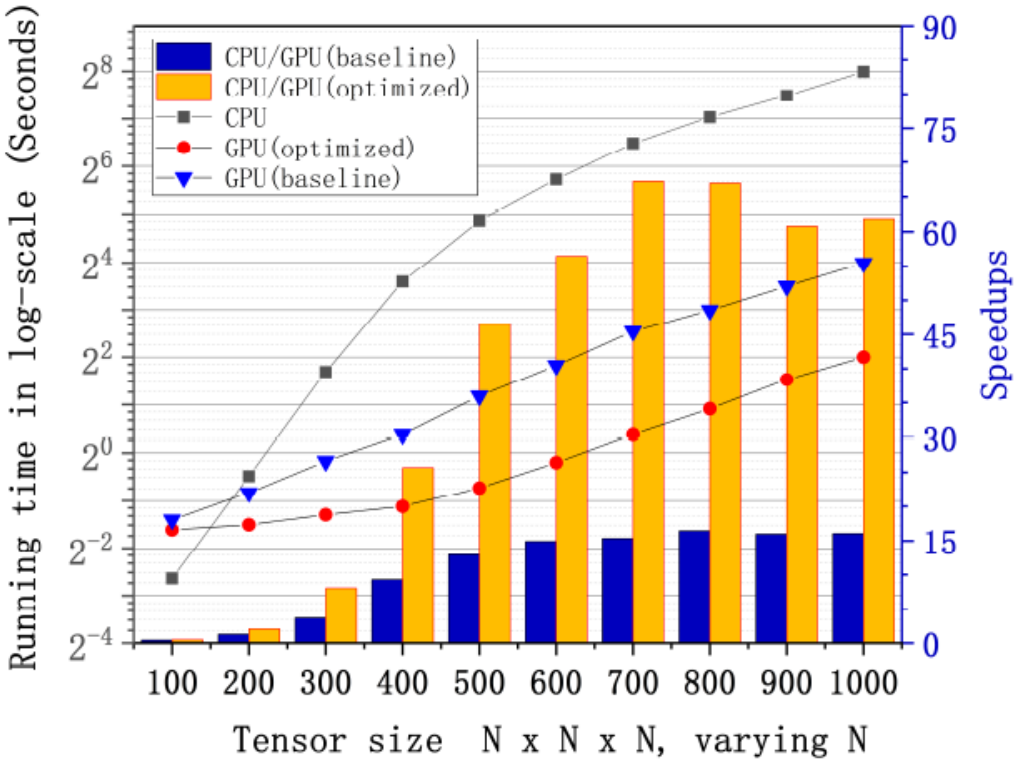
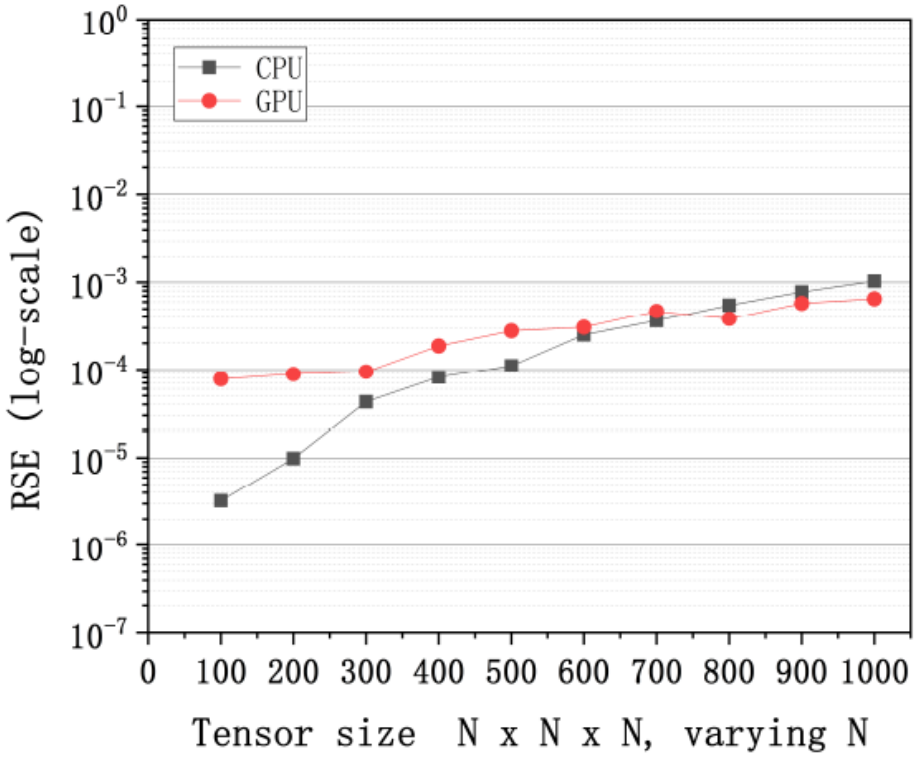
Input: Tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, rank r_1, r_2, r_3, r_4.

1: **for** $i = 1$ to 3 **do**
2: $\quad \boldsymbol{H}_i = \boldsymbol{X}_{(i)} \boldsymbol{X}_{(i)}^T$
3: **end for**
4: **for** $j = 1$ to 3 in parallel **do**
5: $\quad \boldsymbol{U}_j \leftarrow$($r_j$ leading eigen vectors) of $\boldsymbol{H}_j$
6: **end for**
7: $\boldsymbol{U}_4 \leftarrow$($r_4$ leading left singular vectors) of $\boldsymbol{X}_{(3)}^T$
8: $\mathcal{B}_2 = \mathcal{U}_4 \times_1 \boldsymbol{U}_1^T \times_2 \boldsymbol{U}_2^T$,
9: $\mathcal{B}_1 = \mathcal{X} \times_1 \boldsymbol{U}_4^T \times_2 \boldsymbol{U}_3^T$,

Output: $\mathcal{B}_1, \mathcal{B}_2, \boldsymbol{U}_3, \boldsymbol{U}_2, \boldsymbol{U}_1$.

Design and Implementation

- Using a one-dimensional array of data and stored in columns in the GPU, the matricization operation can be omitted.
- Take the batch solution to calculate intermediate variables in parallel.
- Use shared memory to store data, which has low latency and high bandwidth relative to global memory.



Experiment Results