cuTensor-CP: High Performance Third-order CP Tensor Decomposition on GPUs

Xiao-Yang Liu, Han Lu, Tao Zhang

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

- Tensor decompositions have become a powerful tool for big data analytics and machine learning.
- Time and space complexities of tensor decomposition algorithms grow rapidly with the size of tensors.
- Exploiting parallelisms of tensor algorithms and accelerating them on many-core GPUs are promising.

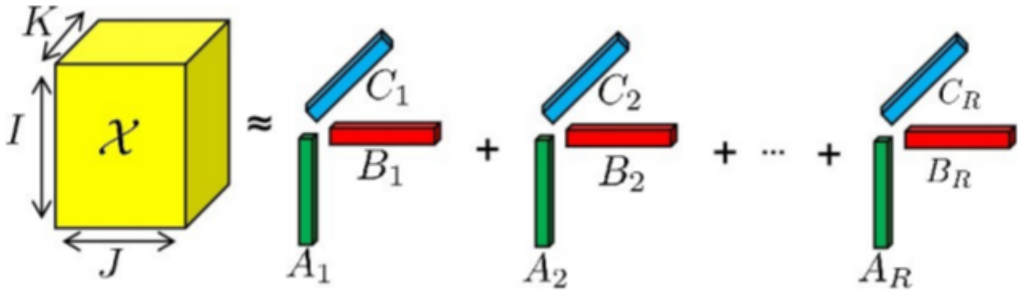
Contributions

We implement key tensor operations, including tensor matricization and matricized tensor times Khatri-Rao product (MTTKRP).

We implement and optimize whole CP tensor decomposition on GPUs.

We perform numerical experiments to evaluate the performance of MTTKRP and CP tensor decomposition.

Parallel CP Decomposition on the GPU



The CP tensor decomposition factorizes a tensor into the sum of rank-one tensor components.

Alternating least square

Algorithm 1 ALS CP tensor decomposition

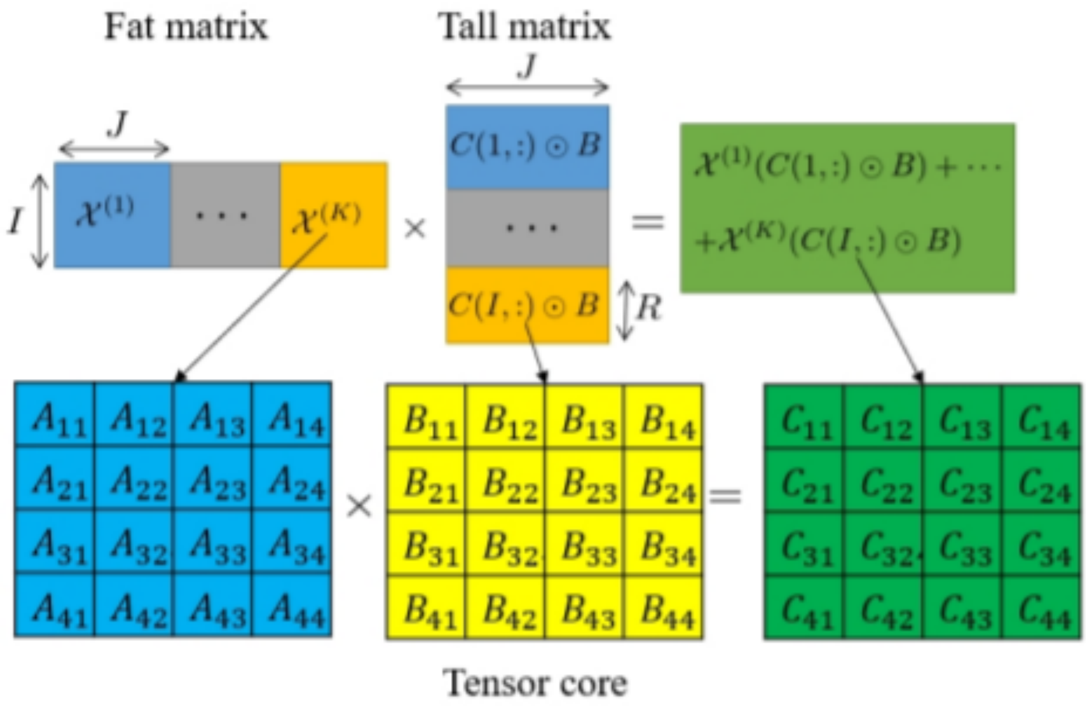
1: **Input**: tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$, rank R.
2: Randomly initialize $\boldsymbol{A} \in \mathbb{R}^{I \times R}, \boldsymbol{B} \in \mathbb{R}^{J \times R}, \boldsymbol{C} \in \mathbb{R}^{K \times R}$,
3: **while** convergence criterion is not met **do**
4: $\quad \boldsymbol{A} \leftarrow \boldsymbol{X}_{(1)}(\boldsymbol{C} \odot \boldsymbol{B})(\boldsymbol{C}^\top \boldsymbol{C} .* \boldsymbol{B}^\top \boldsymbol{B})^\dagger$,
5: $\quad \boldsymbol{B} \leftarrow \boldsymbol{X}_{(2)}(\boldsymbol{C} \odot \boldsymbol{A})(\boldsymbol{C}^\top \boldsymbol{C} .* \boldsymbol{A}^\top \boldsymbol{A})^\dagger$,
6: $\quad \boldsymbol{C} \leftarrow \boldsymbol{X}_{(3)}(\boldsymbol{B} \odot \boldsymbol{A})(\boldsymbol{B}^\top \boldsymbol{B} .* \boldsymbol{A}^\top \boldsymbol{A})^\dagger$,
7: **end while**
8: **Output**: $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$.

Lines 3-7 are the iterative process and the algorithm updates the factor matrices (lines 4-6) alternatively.

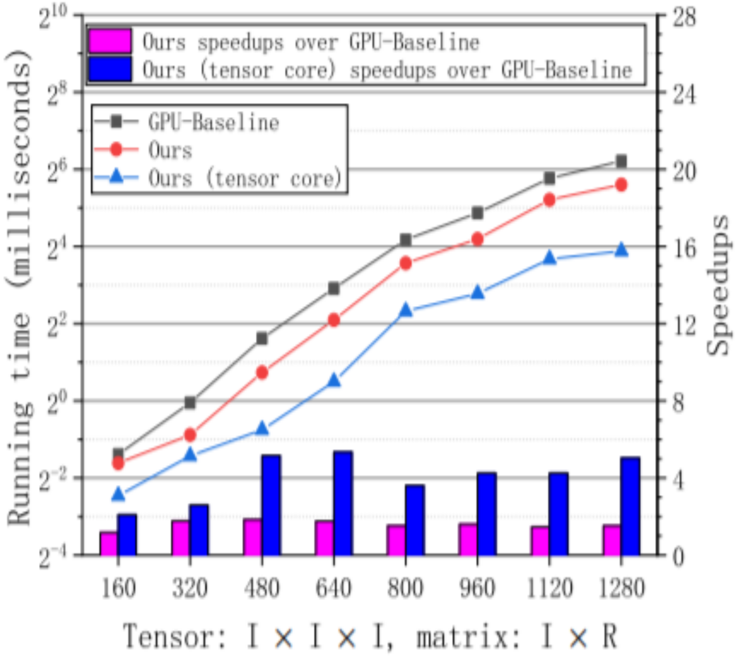
Design and Implementation

- Tensor matricization results in a fat matrix and the Khatri-Rao product results in a tall matrix

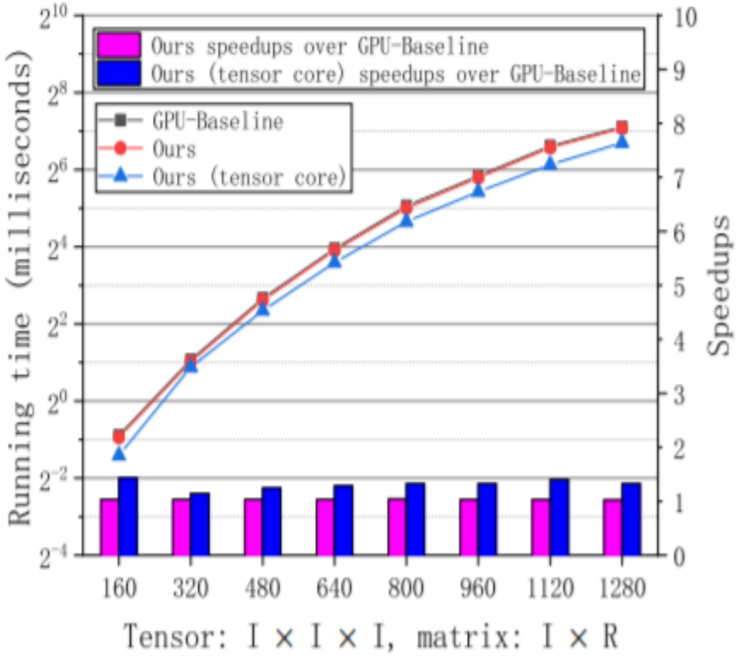
- Matrix multiplication can be calculated in a block manner, we divide the large matrices into smaller matrices and batch the block matrix multiplications onto tensor cores.



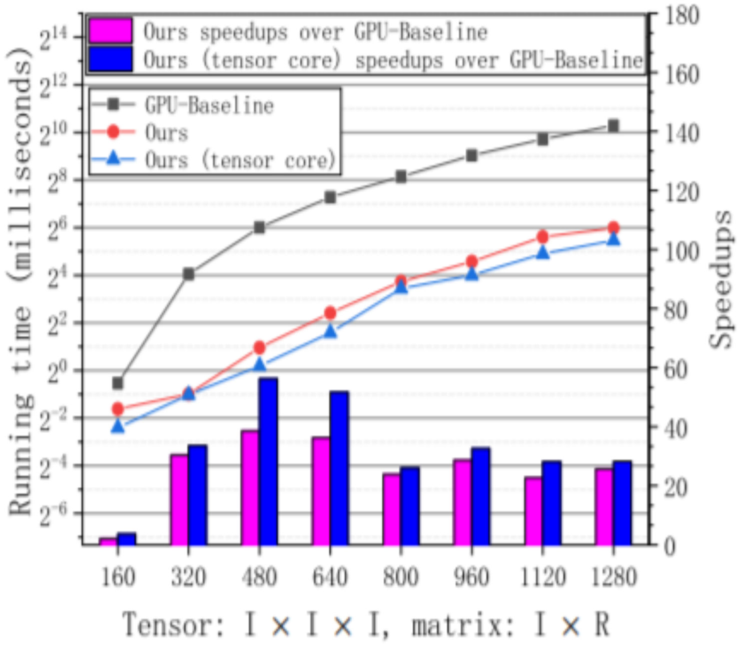
Experiment Results



(a) mode-1 MTTKRP.

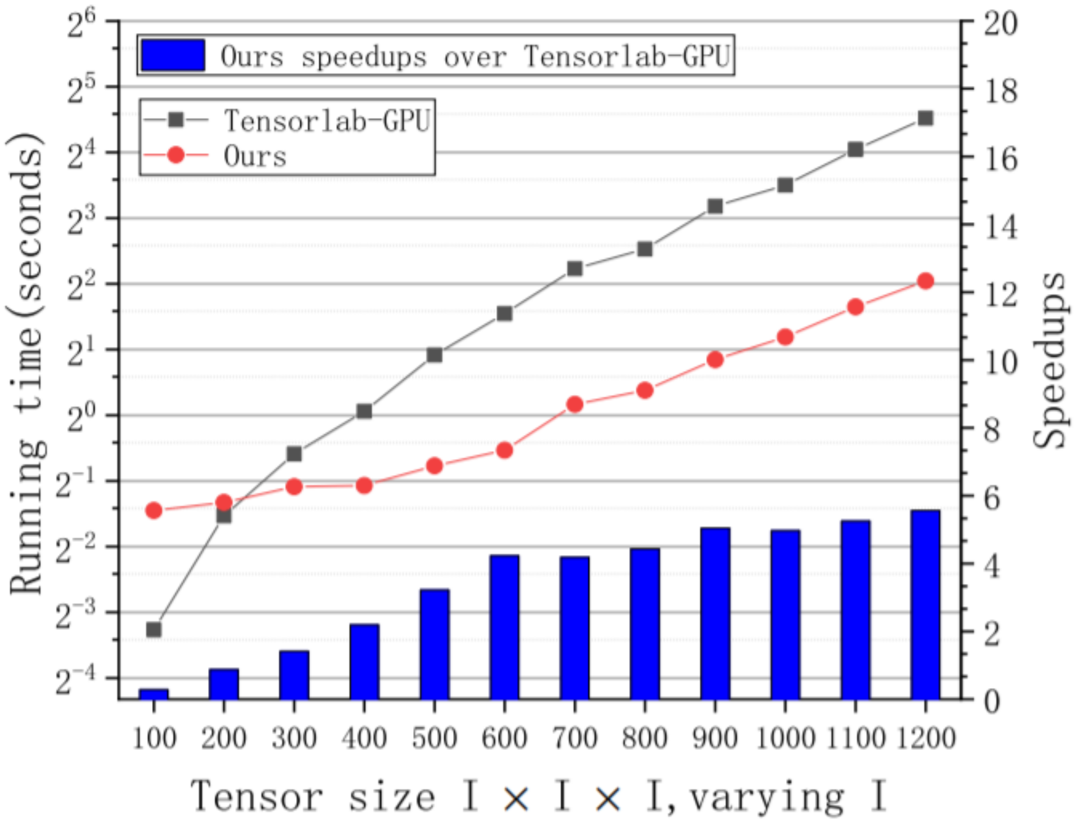


(b) mode-2 MTTKRP.



(c) mode-3 MTTKRP.

Experiment Results



Our GPU implementation achieves up to 5.56× speedup versus the TensorLab-GPU.