cuTensor-TT/TR: High Performance Third-order Tensor-Train and Tensor-Ring Decompositions on GPU

TNRML Workshop, IJCAI 2020

Hao Hong, Tao Zhang, Xiao-Yang Liu
Shanghai University, Columbia University

Motivations

- Tensor decompositions have become basic tools in many fields.
  - **Data mining**
  - **Computer vision**
  - **Deep learning**

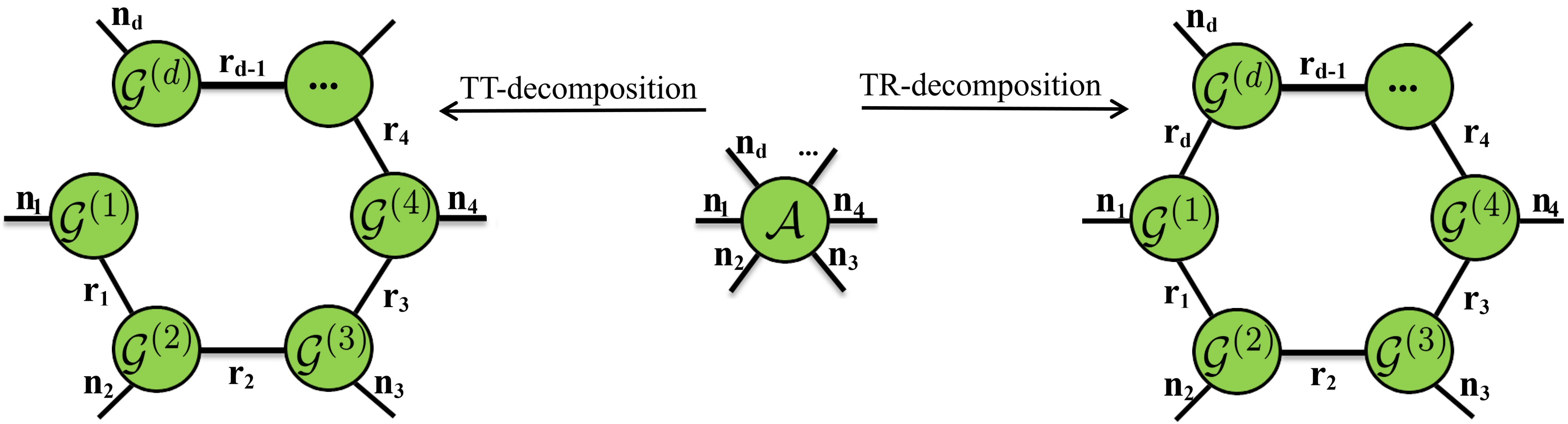
- With the ever-growing demands of efficient big data analytics, developing efficient tensor decompositions becomes a critical task.
  - **Dimensionality increase**
  - **Order increase**

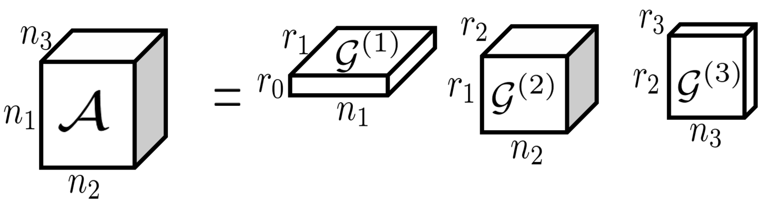
- The existing optimizations are not incompatible with the GPU architecture.
  - **Optimize better to the GPU architecture**

Contributions

- Implement third-order TT and TR decompositions on GPU
- Optimizing memory access, parallel strategies, faster data transfer, and faster tensor products
- Achieve up to 6.67× and 6.36× speedup respectively vs. the GPU-baseline on GPU

TT and TR Decompositions





$$\mathcal{A} = \mathcal{G}^{(1)} \circ \mathcal{G}^{(2)} \circ \mathcal{G}^{(3)}$$

Optimization Strategies

- **Faster Memory Access**
  - Tensor $\mathcal{X}$ in a 1D array $\boldsymbol{x}$
- **Parallelization Schemes**
  - Parallel Jacobi SVD
  - Parallel diagonal matrix times matrix

$$SV^T = \text{parallel}(s_k \cdot V_k^T)$$

  - Parallel element-wise product

$$\text{parallel}(s_{m-k+1}^{(0)} = s_k \cdot s_k),\ 1 \leq k \leq m$$

- **Data Transfer**
  - Streaming transmission modules

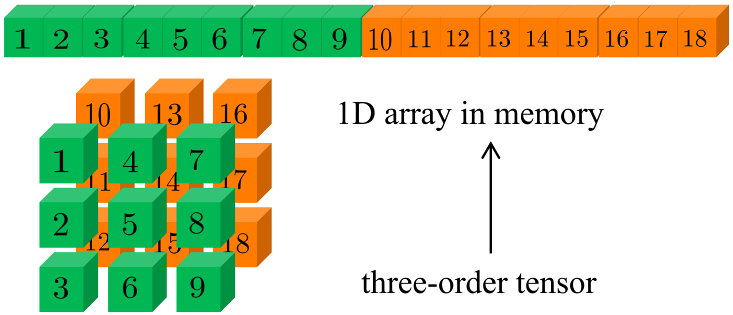


Figure 3: Tensor's storage as a 1D array in GPU memory

Performance on GPU

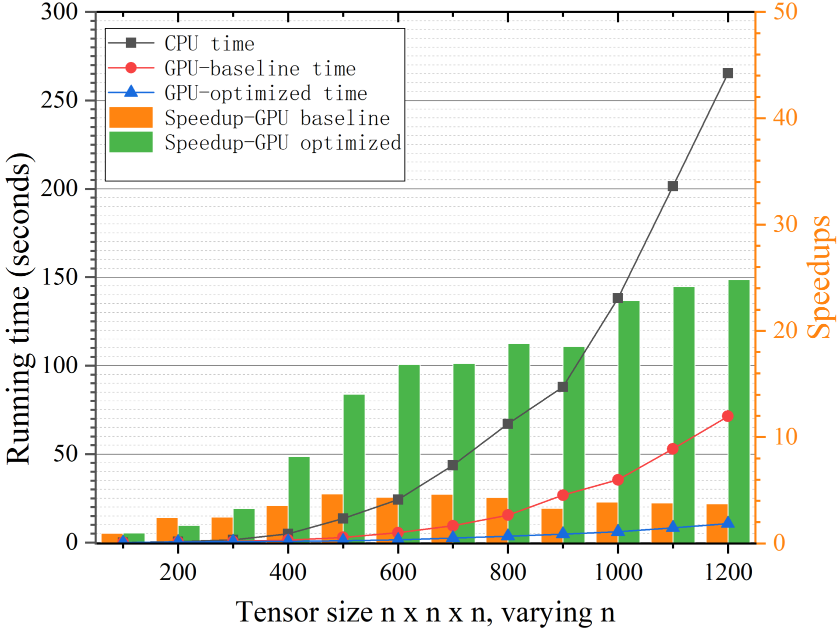


Figure 4: Running time and speedups of third-order TT decomposi-tion on Tesla V100 GPU and two 10-core CPUs

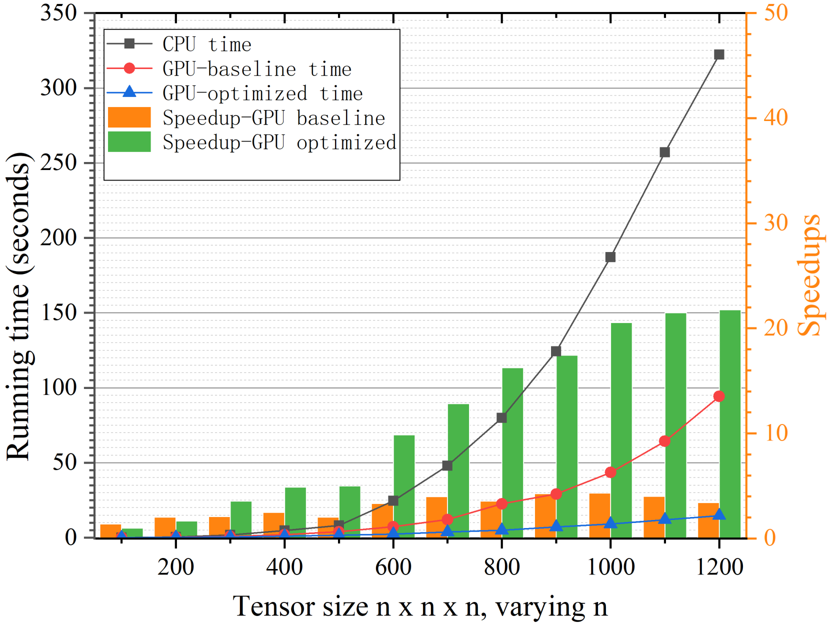


Figure 5: Running time and speedups of third-order TR decomposi-tion on Tesla V100 GPU and two 10-core CPUs.

Conclusion

- According to the characteristics of the algorithms and the architecture of GPU, we implemened **third-order tensor-train and tensor-ring decompositions on GPU, exploiting parallelism.**

- We designed **three optimization strategies:** parallelization schemes, optimizing memory access , etc. We achieved **up to 6.67× and 6.36×** speedups for third-order TT and TR decompositions over the GPU-baseline on a Tesla V100 GPU.

- Future work: higher-order decomposition, multi-node GPU implementation, incorporating TT and TR decompositions into the cuTensor library.

Thank you!