Compact Decomposition of Irregular Tensors for Data Compression: From Sparse to Dense to High-order Tensors

Taehyung Kwon[1], Jihoon Ko[1], Jinhong Jung[2], Jun-Gi Jang[3], Kijung Shin[1]
[1]*KAIST: {taehyung.kwon, jihoonko, kijungs}@kaist.ac.kr,*
[2]*Soongsil Univ: {jinhong}@ssu.ac.kr,* [3]*UIUC: {jungi}@Illinois.edu*

KDD2024 BARCELONA, SPAIN

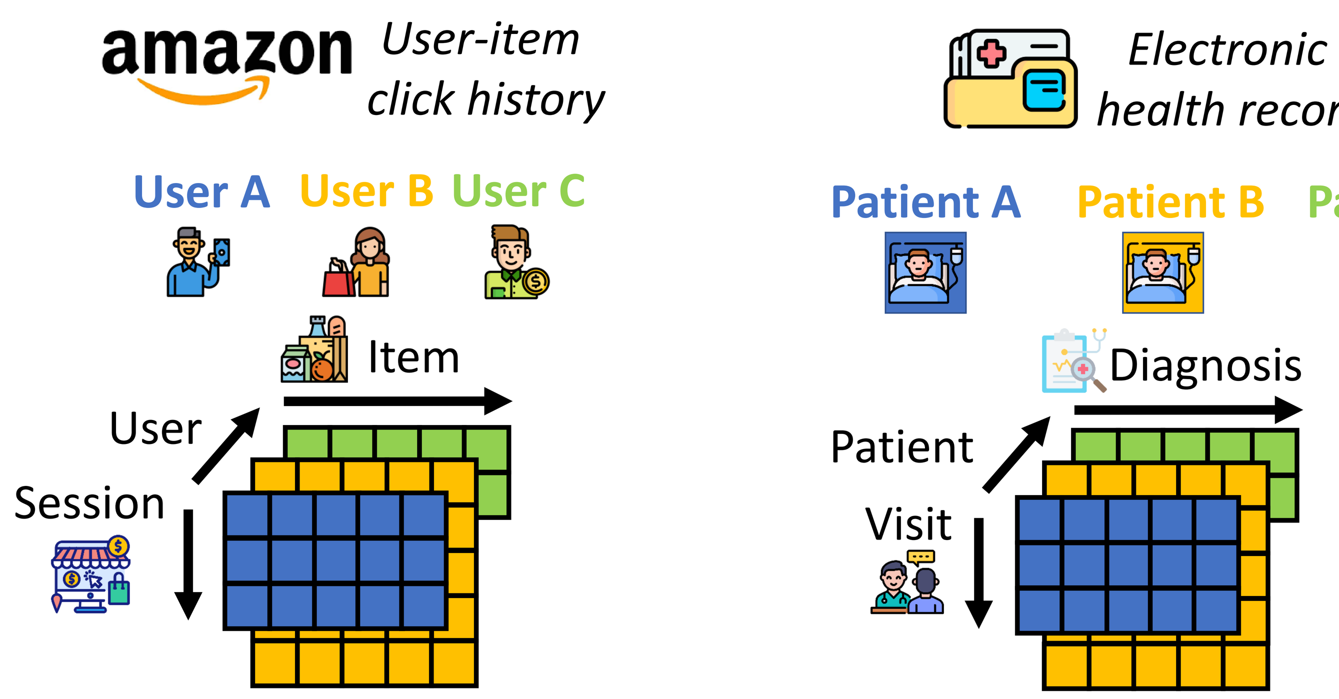
GitHub

Summary

- **Goal:** We developed new **decompositions for irregular tensors** that are specialized in **lossy compression**.
- **Advantage 1) Compact Compression:** The compressed outputs of our methods are up to 37x smaller than those of the most concise baseline.
- **Advantage 2) Accurate Decompression:** Our methods are up to 5x more accurate than the most accurate baseline.
- **Advantage 3) Versatile Application:** Our methods are effective for sparse, dense, and higher-order irregular tensors.

Introduction

- **Irregular tensor:** A (3-order) irregular tensor is a collection of matrices with varying row counts.

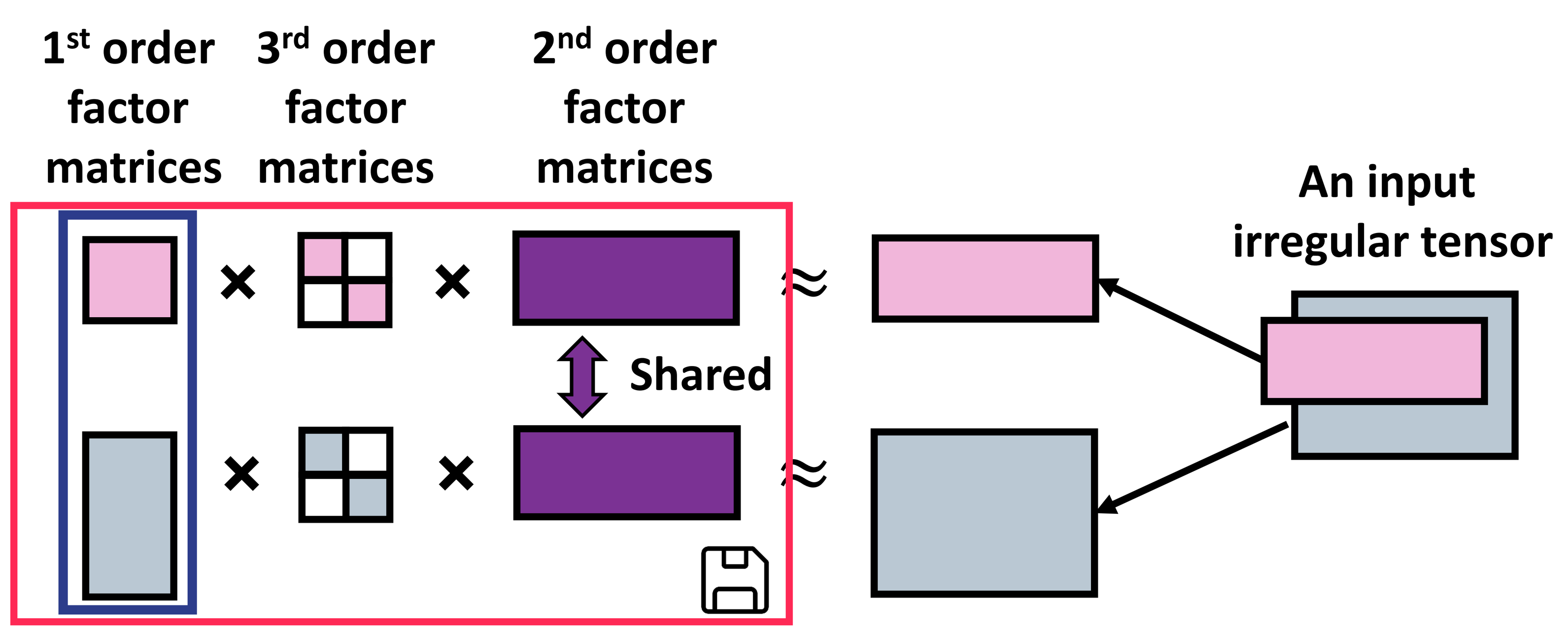


Examples of real-world irregular tensors
⚠ More than 140 billion entries

- **Research question:** How can we accurately and compactly compress irregular tensors of any order and density?

PARAFAC2 and its limitation

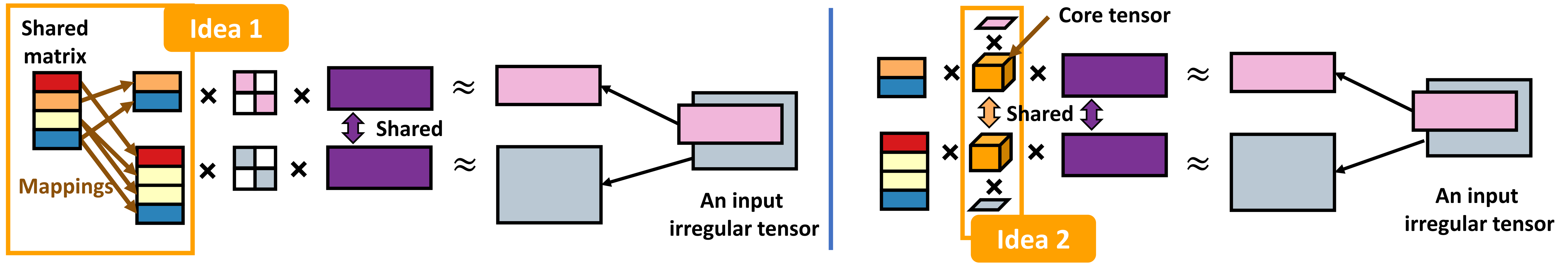
- **PARAFAC2 [1]:** A compression method for irregular tensors.



- **Memory bottleneck of PARAFAC2:** 1st mode factor matrix is required per slice. ➡ accounts for majority of compressed size.

Proposed method: Light-IT and Light-IT++

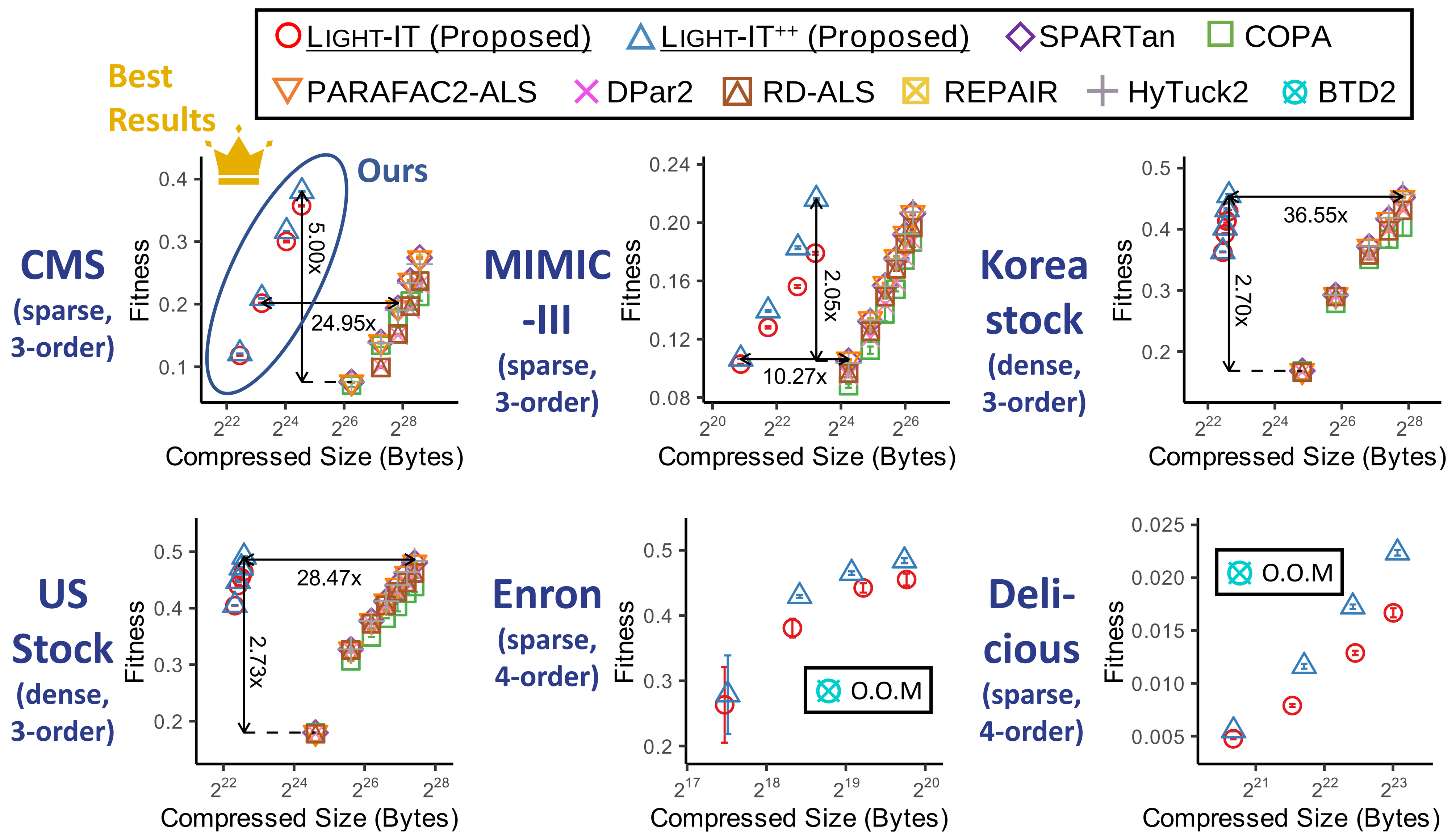
- **Improved conciseness (Light-IT):** Use a single **shared factor matrix** instead of multiple different 1st mode factor matrices. Each row of each 1st mode factor matrix corresponds to one of the rows in the shared matrix.
- **Improved approximation (Light-IT++):** Incorporate a **core tensor** for better approximation as in the Tucker model [3]. The core tensor is trained to capture relationships between different latent features.



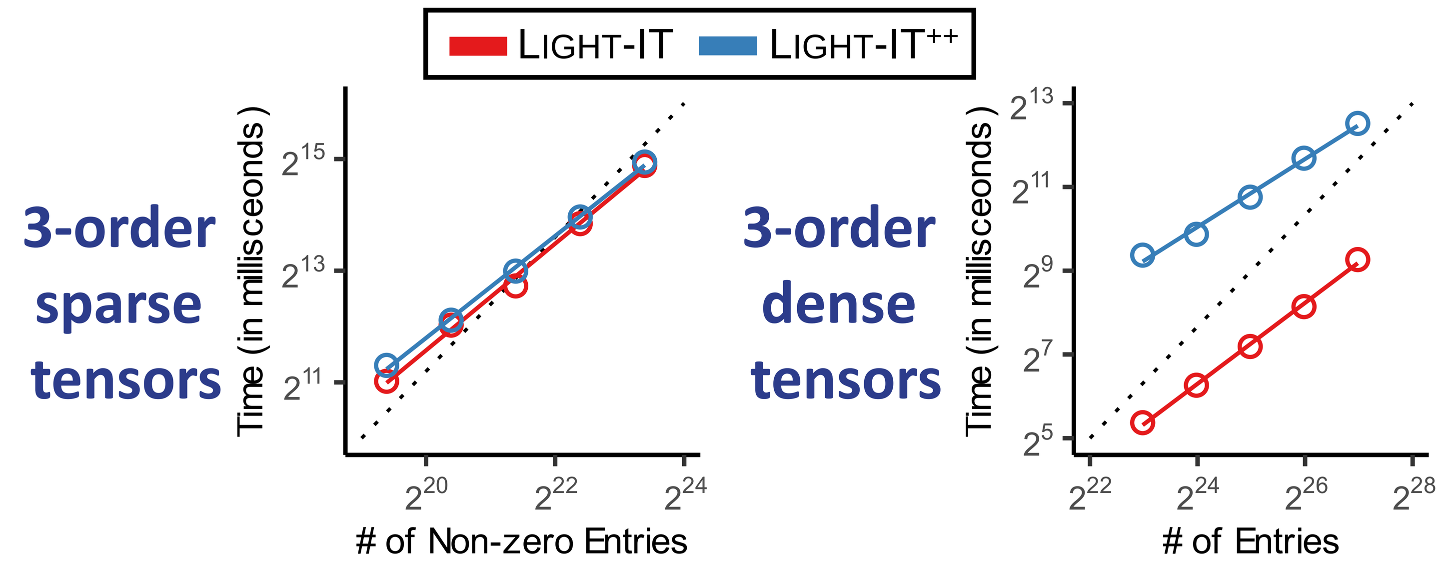
- **Sparse design:** Exploit sparsity of **sparse input tensors** for speed by computing the loss for all zero entries efficiently in a closed form.
- **Higher-order design:** Applied to **input tensors of any order** by matricizing the input irregular tensor and its approximation.

Experiments

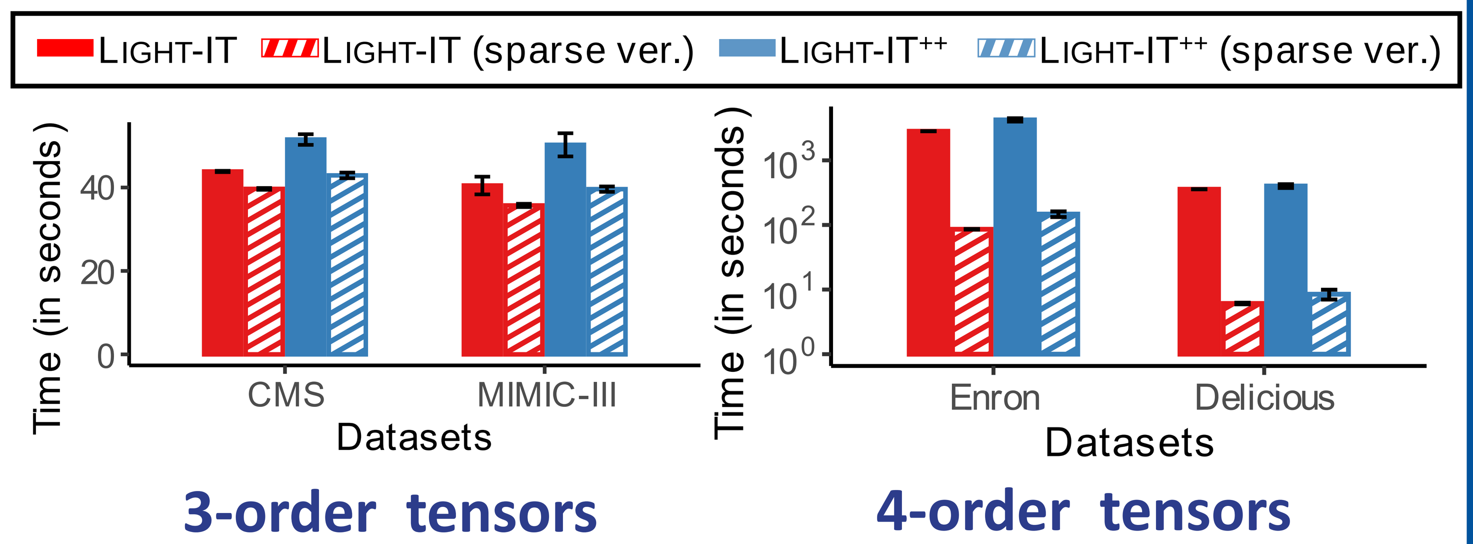
- **Compression performance:** Light-IT and Light-IT++ are **compact** and **accurate**.
  - **Datasets:** 6 real-world irregular tensors (e.g., EHR and stock data), including sparse and dense tensors of order 3 and 4.
  - **Baselines:** 8 state-of-the-art methods for irregular tensor decomposition.



- **Scalability:** Our methods **scale near-linearly** with the number of (non-zero) entries of a (sparse) tensor.



- **Efficiency in sparse tensors:** For sparse irregular tensors, the sparse versions of our methods are faster than their dense versions.



Reference

[1] Richard A Harshman et al. 1972. PARAFAC2: Mathematical and technical notes. UCLA working papers in phonetics 22, 3044 (1972), 122215.
[2] Ting Chen, Lala Li, and Yizhou Sun. 2020. Differentiable product quantization for end-to-end embedding compression. In ICML.
[3] Ledyard R Tucker. 1966. Some mathematical notes on three-mode factor analysis. Psychometrika 31, 3 (1966), 279–311.