Tensor Network Representations in Machine Learning

Qibin Zhao

Tensor Learning Unit
RIKEN AIP

July 31, 2019

Research Goal

To advance machine learning methods by leveraging tensor network representations

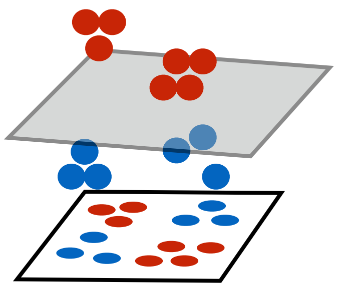
- Model compression
- Tensor completion
- Multi-task learning
- Multi-modal learning

Background & Motivation

Machine Learning

$$f(\mathbf{x}) = W \cdot \Phi(\mathbf{x})$$

- Curse of dimensionality
- Nonlinear mapping is unknown



Kernel Learning

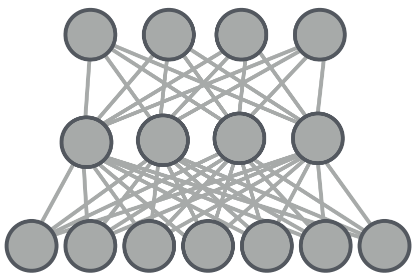
Kernel learning

$$W = \sum_j \alpha_j \Phi(x_j)$$

- "kernelization" scales quadratically with training data size
- Low generalization due to representer theorem

Deep neural network

$$f(\mathbf{x}) = \Phi_2\Big(M_2\Phi_1(M_1\mathbf{x})\Big)$$



Neural Nets

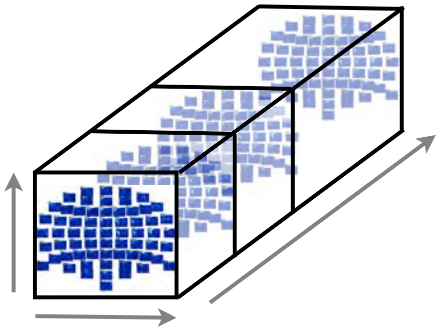
- Model parameters are huge (space)
- Computational inefficient due to model complexity (time)

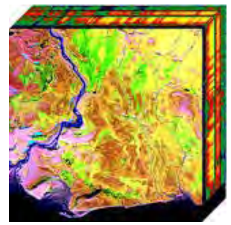
Are tensor networks useful in solving these problems ?

Background & Motivation

High-order structured data

- Video, Hyperspectral image, fMRI, EEG
- Social network (user x user x relation)
- …





Multi-modal, multi-view learning

- Multi-linear mapping: $f(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) = \mathcal{W} \times_1 \Phi(\mathbf{x}_1) \times_2 \Phi(\mathbf{x}_2) \times_3 \Phi(\mathbf{x}_3)$

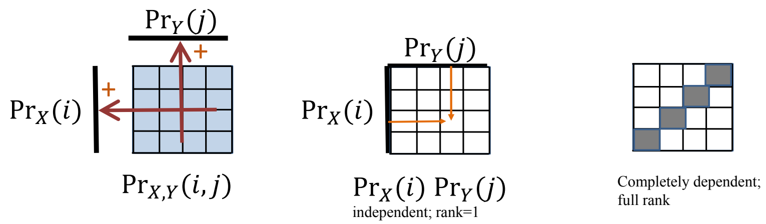
Multi-task deep learning

- Model parameters $\{W_n, \forall n = 1, \ldots, T\}$ form a tensor.

High-order moment, joint PMF

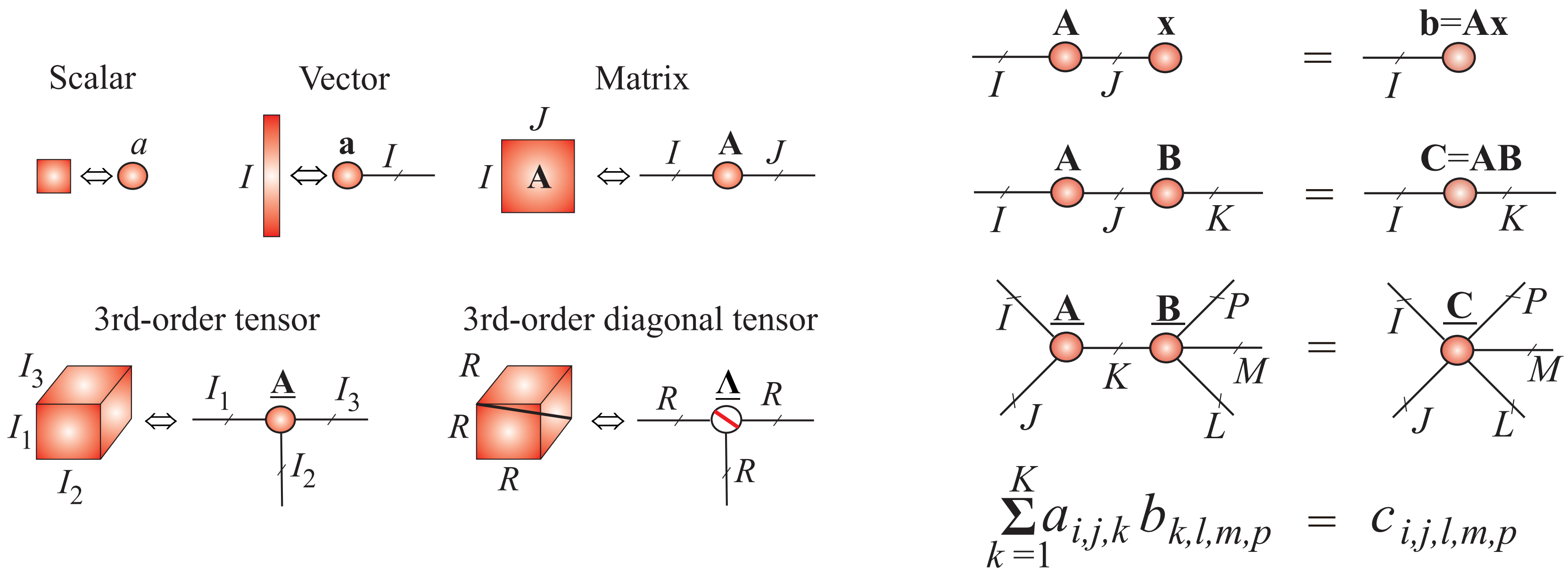
$\mathbb{E}[x \otimes x \otimes x] \in \mathbb{R}^{d \times d \times d}$ is a third order tensor.

$\mathbb{E}[x \otimes x \otimes x]_{i_1, i_2, i_3} = \mathbb{E}[x_{i_1} x_{i_2} x_{i_3}]$.



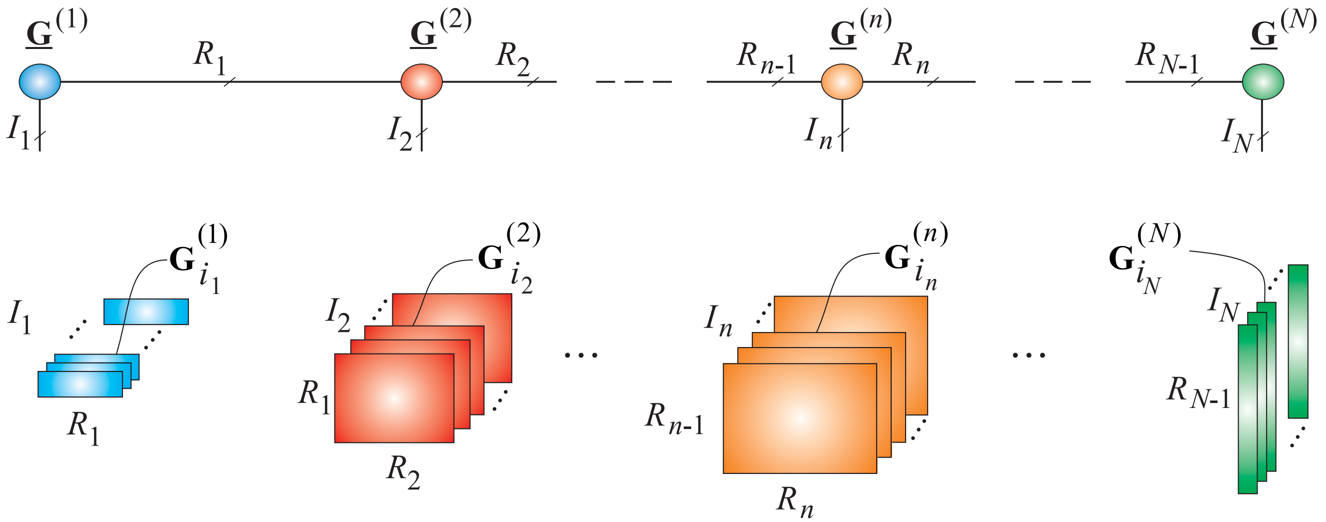
What Are Tensor Networks (TNs) ?

- A powerful tool to describe strongly entangled quantum many-body systems in physics
- Decompose a high-order tensor into a collection of low-order tensors connected according to a network pattern
- Tensor network diagram



TT/MPS Representation and Properties

[V. Oseledets, SIAM J. Sci. Comput., 2011]



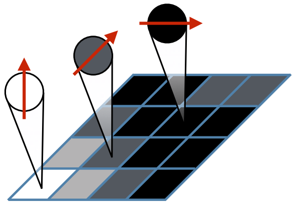
TT: tensor train decomposition; MPS: matrix product state

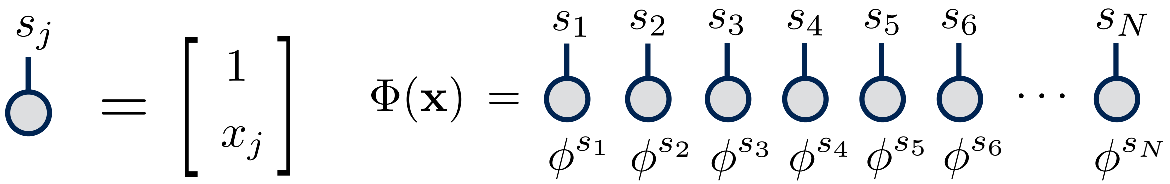
- Efficient to represent I^N data values by $\mathcal{O}(NIR^2)$ parameters
- Efficient to compute or optimize TT/MPS by DMRG algorithm

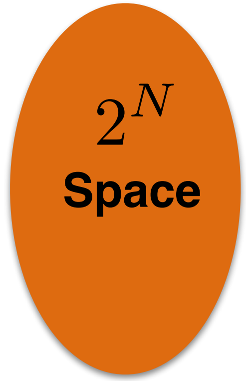
TNs for Weight Compression & Kernel Learning

[E. Stoudenmire, NIPS 2016]

- Input: $\mathbf{x} = [x_1, \; x_2, \; x_3, \; \ldots \;, \; x_N]$
- Nonlinear mapping by tensor product (Hilbert space)

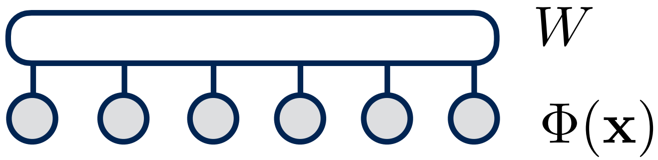






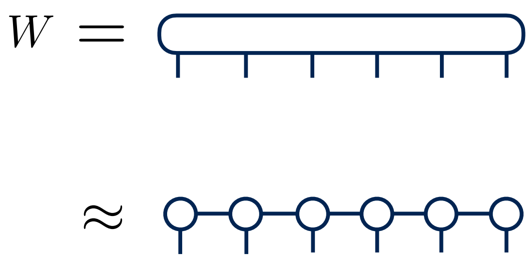
- Decision function - W is an *Nth*-order tensor

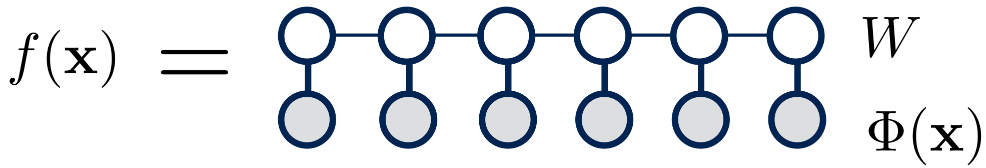
$$f(\mathbf{x}) = W \cdot \Phi(\mathbf{x}) =$$



- TT representation of weight parameter

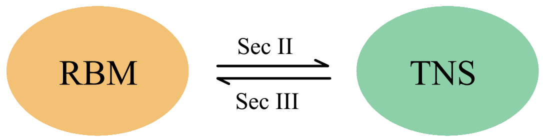
[A. Novikov, NIPS 2015]





Relations Between TNs and DNNs

- Equivalence of Restricted Boltzmann Machines and Tensor Networks

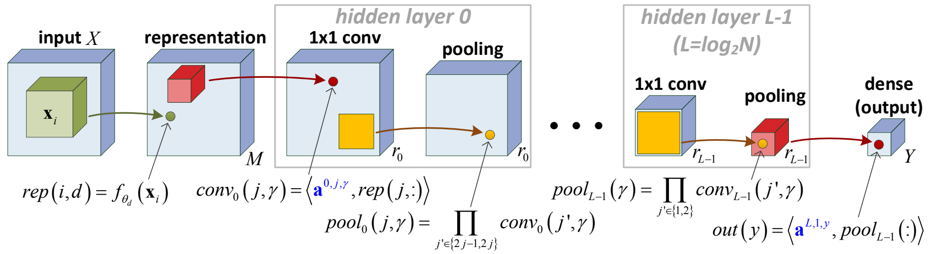


[Chen et al, Physical Review B, 2018]

[Carleo et al, Science, 2017]

- Equivalence of Deep Convolutional Network and Hierarchical Tucker

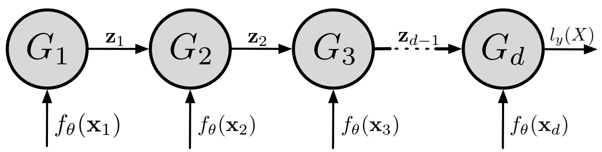
[N. Cohen & A. Shashua, ICML 2016]



network structure (depth, width, pooling etc) $\longleftrightarrow$ decomposition type (dim tree, internal ranks etc)

network weights $\longleftrightarrow$ decomposition parameters

- Recurrent Neural Networks and Tensor Train *[Khrulkov, ICLR 2018]*

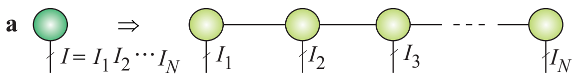


Tensor Decompositions	**Deep Learning**
CP-decomposition	shallow network
TT-decomposition	RNN
HT-decomposition	CNN
rank of the decomposition	width of the network

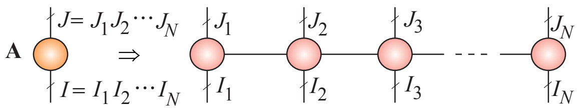
- Powerful tools to study theory behind DNN

Tensor Networks for Large-Scale Optimization Problems

- TT format of a large vector



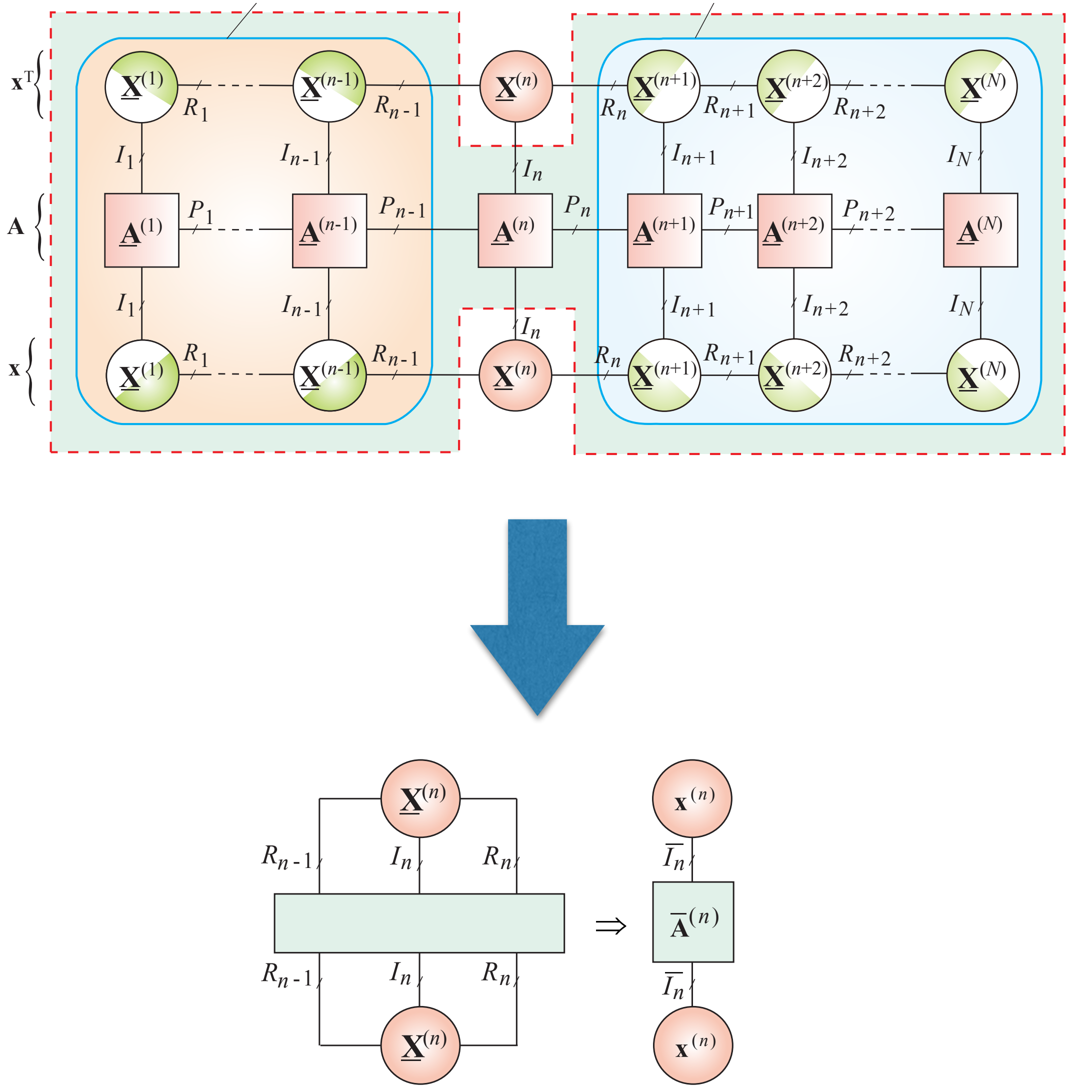
- TT format of a large matrix



- Fast ALS/DMRG algorithm

- Applicable to large-scale SVD/PCA/CCA and etc

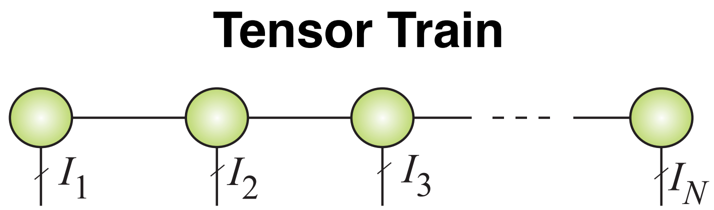
Eigenvalue problem: $\max x^T A x$



Fundamental Tensor Network Model

Tensor train (TT) representation

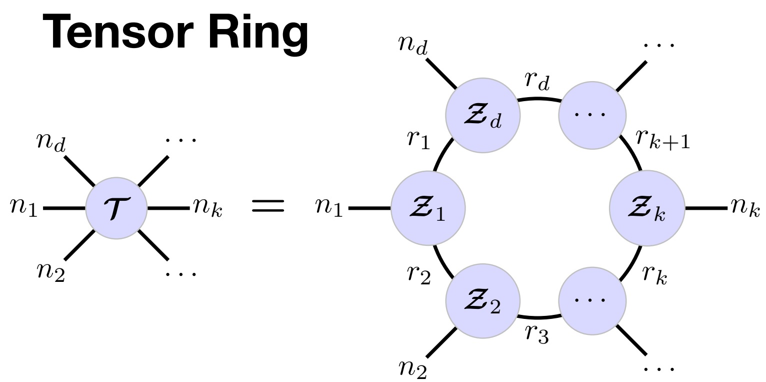
- Powerful but still some limitations
- TT-ranks of middle cores are large



Tensor ring representation

- Generalized TT without constraints on boundary cores
- Efficient computation for multilinear operations
- Highly expressive capacity

$$x_{i_1,i_2,\ldots,i_N} = \mathrm{tr}\,(\mathbf{G}^{(1)}_{i_1}\ \mathbf{G}^{(2)}_{i_2}\ \cdots\ \mathbf{G}^{(N)}_{i_N})$$



[Zhao et al, ICLR workshop 2018, ICASSP 2019]

Efficient Operations

- Sum of tensors

$$\mathcal{T}_1 = \Re(\mathcal{Z}_1, \ldots, \mathcal{Z}_d) \quad \mathcal{T}_3 = \mathcal{T}_1 + \mathcal{T}_2,$$
$$\mathcal{T}_2 = \Re(\mathcal{Y}_1, \ldots, \mathcal{Y}_d), \quad \mathcal{T}_3 = \Re(\mathcal{X}_1, \ldots, \mathcal{X}_d),$$

$$\mathbf{X}_k(i_k) = \begin{pmatrix} \mathbf{Z}_k(i_k) & 0 \\ 0 & \mathbf{Y}_k(i_k) \end{pmatrix}, \quad \begin{matrix} i_k = 1, \ldots, n_k, \\ k = 1, \ldots, d. \end{matrix}$$

- Multilinear products

$$\mathcal{T} = \Re(\mathcal{Z}_1, \ldots, \mathcal{Z}_d) \qquad c = \mathcal{T} \times_1 \mathbf{u}_1^T \times_2 \cdots \times_d \mathbf{u}_d^T$$

$$c = \Re(\mathbf{X}_1, \ldots, \mathbf{X}_d) \textit{ where } \mathbf{X}_k = \sum_{i_k=1}^{n_k} \mathbf{Z}_k(i_k) u_k(i_k).$$

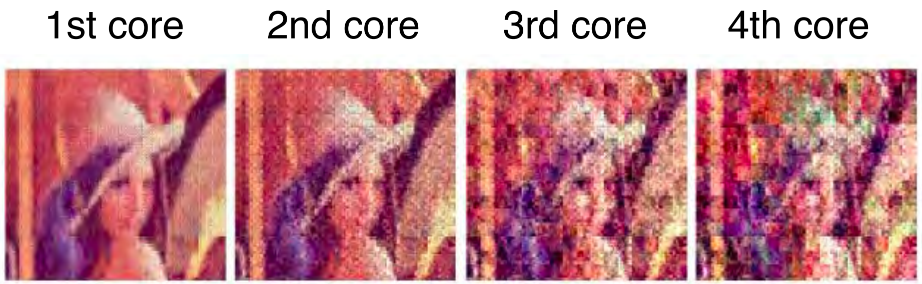
- Hadamard product of tensors

$$\mathcal{T}_3 = \mathcal{T}_1 \circledast \mathcal{T}_2 = \Re(\mathcal{X}_1, \ldots, \mathcal{X}_d), \qquad \mathbf{X}_k(i_k) = \mathbf{Z}_k(i_k) \otimes \mathbf{Y}_k(i_k), \quad k = 1, \ldots, d.$$

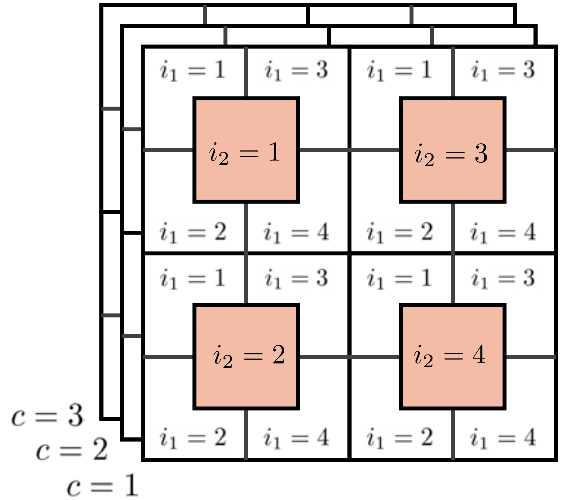
- Inner product of two tensors
  - Apply Hadamard product followed by multilinear products with vectors of all ones.

Tensor Ring Representation

- High-order structure relations can be captured
- Compact representation by many small cores
- Interpretability



[Bengua et al, IEEE TIP, 2017]



Higher-order tensorization

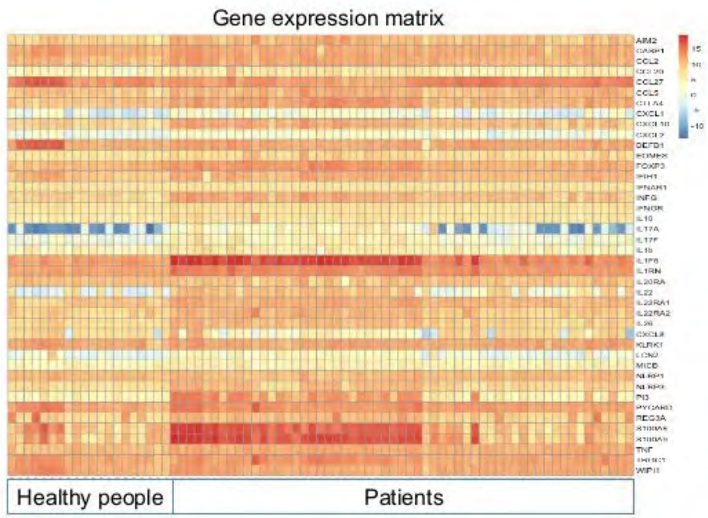
Table 4: Image representation by using tensorization and TR decomposition. The number of parameters is compared for SVD, TT and TR given the same approximation errors.

Data	$\epsilon = 0.1$		$\epsilon = 0.01$		$\epsilon = 9e-4$		$\epsilon = 2e-15$	
$n = 256, d = 2$	SVD	TT/TR	SVD	TT/TR	SVD	TT/TR	SVD	TT/TR
	9.7e3	9.7e3	7.2e4	7.2e4	1.2e5	1.2e5	1.3e5	1.3e5
Tensorization	$\epsilon = 0.1$		$\epsilon = 0.01$		$\epsilon = 2e-3$		$\epsilon = 1e-14$	
	TT	TR	TT	TR	TT	TR	TT	TR
$n = 16, d = 4$	5.1e3	3.8e3	6.8e4	6.4e4	1.0e5	7.3e4	1.3e5	7.4e4
$n = 4, d = 8$	4.8e3	4.3e3	7.8e4	7.8e4	1.1e5	9.8e4	1.3e5	1.0e5
$n = 2, d = 16$	7.4e3	7.4e3	1.0e5	1.0e5	1.5e5	1.5e5	1.7e5	1.7e5

Tensor Networks for Data Representation

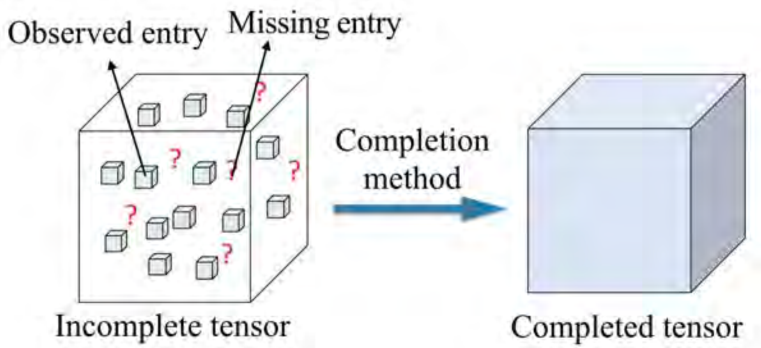
Real data is often high-dimensional

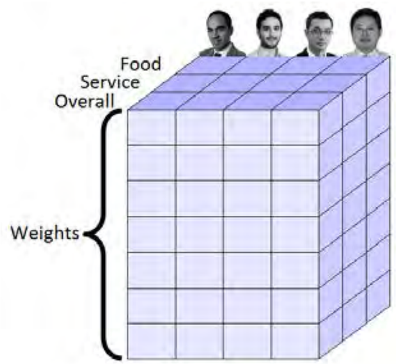
- Recommender system (user x item x time)
- Gene expression, remote sensing, fMRI



Real data is often incomplete

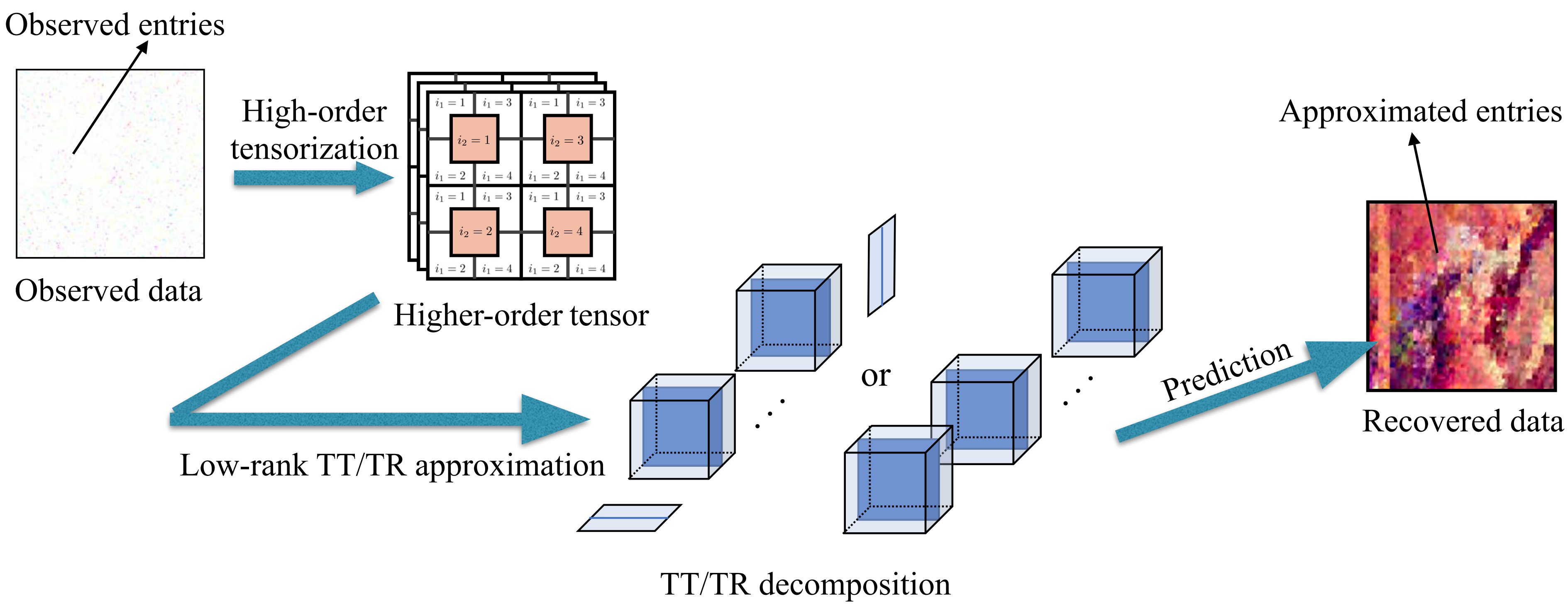
- Low-rank approximation via convex optimization (high computation cost)
- Decomposition based approach (model selection problem)
- How much structure information can be used?





Tensor Networks for Data Imputation

Tensor completion based on TT/TR decomposition



Tensor Ring Low-rank Factors

[Yuan et al, AAAI 2019]

- Tensor ring decomposition with low-rank factors via nuclear norm regularizaiton

Theorem 1. *Given an N-th order tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ which has TR-format, then the following inequality holds for all $n = 1, \ldots, N$:*

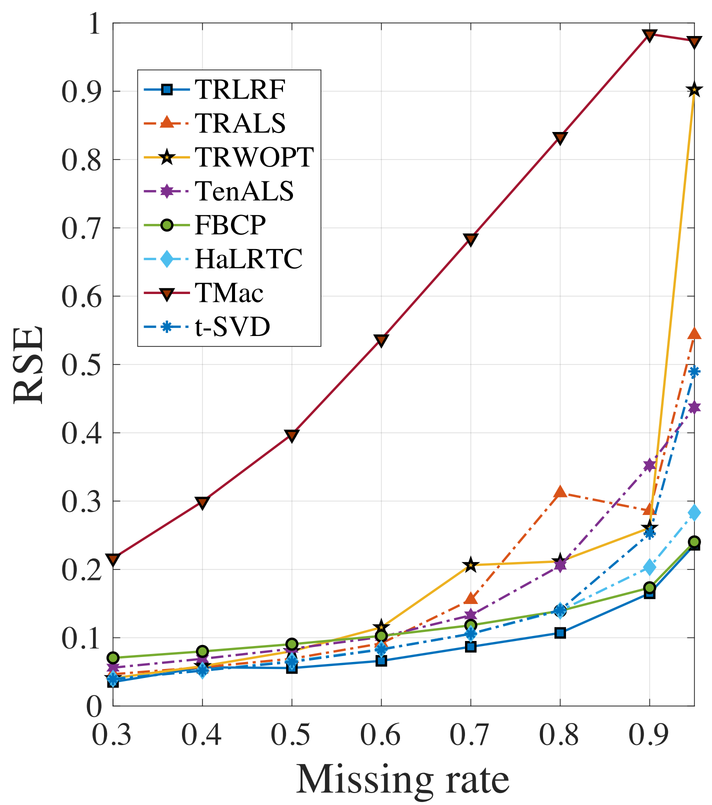
$$Rank(\mathbf{G}^{(n)}_{(2)}) \geq Rank(\mathbf{X}_{(n)}). \tag{1}$$

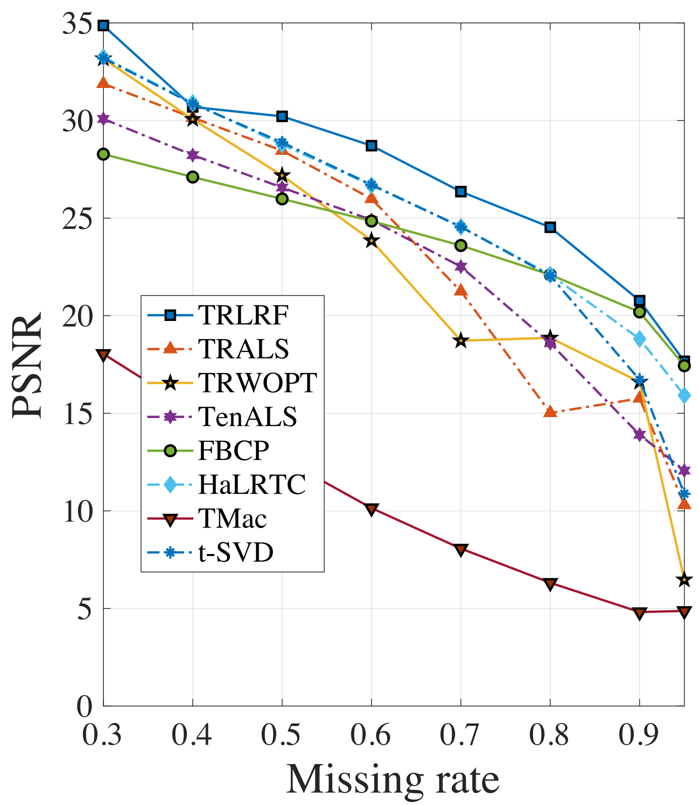
$$\boxed{\sum_{n=1}^{N} \|\mathbf{X}_{(n)}\|_* \longrightarrow \sum_{n=1}^{N} \|\mathbf{G}^{(n)}_{(2)}\|_*}$$

- Theoretically prove the relations between tensor rank and rank of cores
- Robust to rank section by imposing nuclear norm on TR-cores

$$\boxed{\sum_{n=1}^{N} \|\mathbf{G}^{(n)}_{(1)}\|_* + \sum_{n=1}^{N} \|\mathbf{G}^{(n)}_{(3)}\|_*}$$

Experiment Validation





Higher performance than the state-of-the-art algorithms.

TenALS: [Jain, NIPS, 2014]
FBCP: [Zhao, TPAMI, 2015]
HaLRTC: [Liu, TPAMI, 2013]
TMac: [Xu, arXiv, 2013]
t-SVD: [Zhang, CVPR, 2014]

Average performance of 8 benchmark images

Benchmarks

Beyond Unfolding: Reshuffling Operation

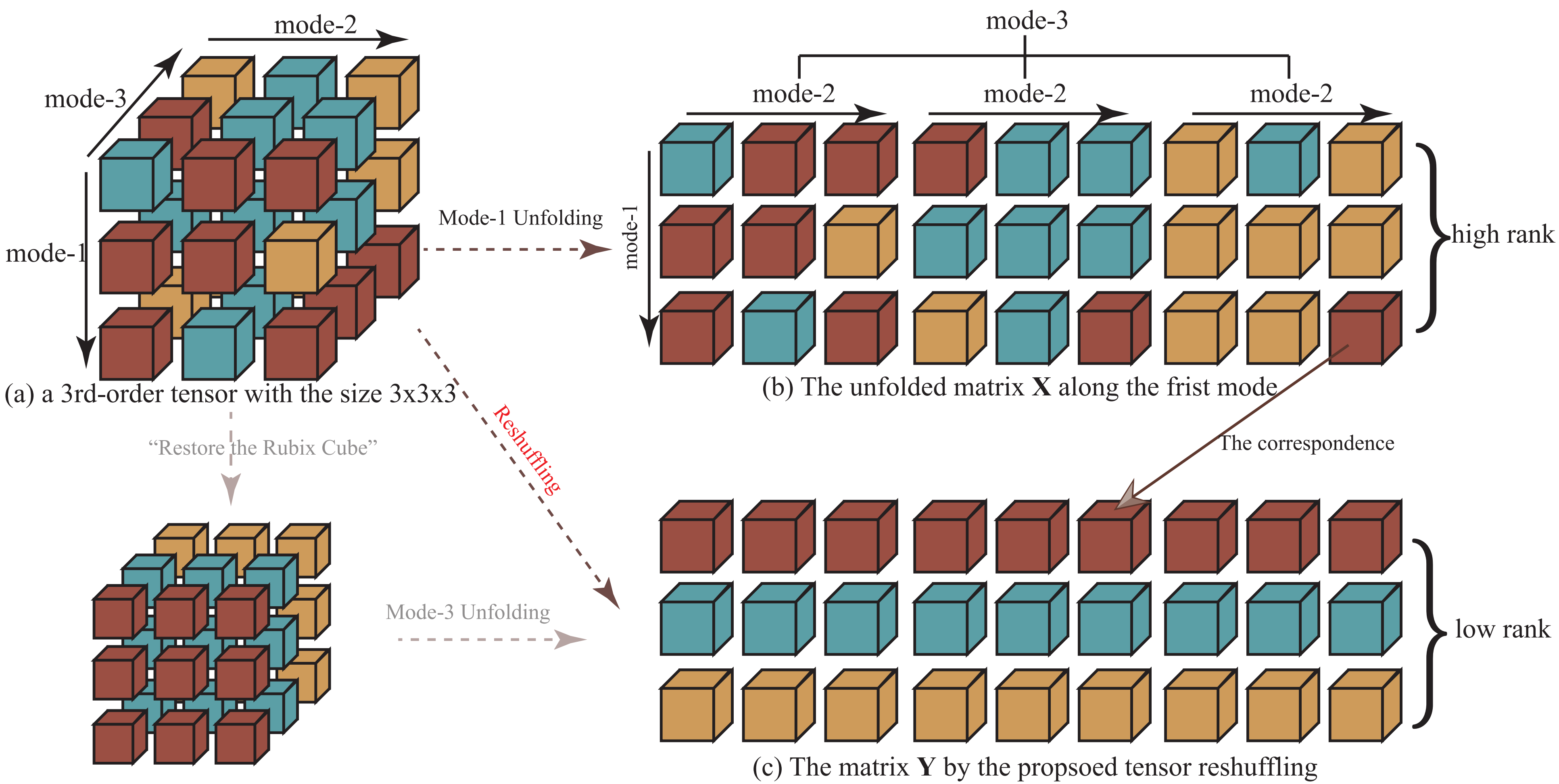
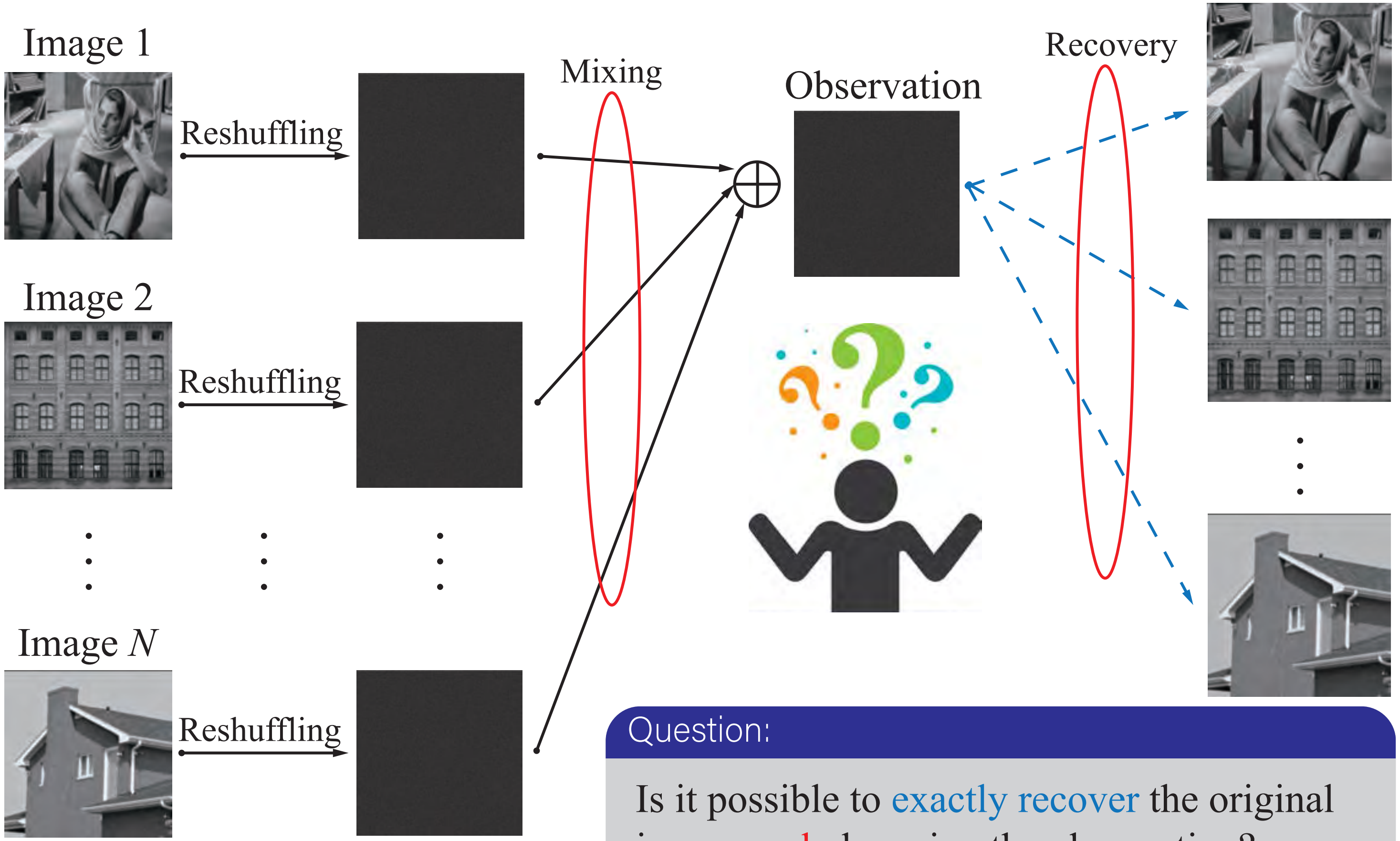


Fig. Difference between tensor unfolding and reshuffling.

Problem Setting



Question:

Is it possible to exactly recover the original images only by using the observation?

Image steganography

Single-shot compressive sensing

Reshuffled Tensor Decomposition

Formulation

Assume that the observation $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_K}$ is a mixture of N components, then the recovery of the original components can be formulated as tensor decomposition, *i.e.*,

$$\mathcal{X} = R_1(\mathbf{A}_1) + R_2(\mathbf{A}_2) + \cdots + R_N(\mathbf{A}_N). \tag{1}$$

where $\mathbf{A}_i$, $i \in [N]$ denote latent components (original images) and R_i denotes the corresponding reshuffling operation *w.r.t.* $\mathbf{A}_i$.

The optimization model:

$$\min_{\mathbf{A}_i,\, i \in [N]} \sum_{i=1}^{N} \|\mathbf{A}_i\|_*, \quad s.t., \mathcal{X} = \sum_{i=1}^{N} R_i(\mathbf{A}_i),$$

where we employ the matrix nuclear norm $\|\cdot\|_*$ in the model as a surrogate of the matrix rank.

Reshuffled Tensor Decomposition

Theoretical results:

Definition: Reshuffled-low-rank incoherence

$$\mu_i(\mathbf{A}) := \max_{j \neq i} \max_{\substack{\mathcal{Y} \in \mathbb{T}_i(\mathbf{A}), \\ \left\| R_i^\star(\mathcal{Y}) \right\|_2 \leq 1}} \left\| R_j^\star(\mathcal{Y}) \right\|_2, \tag{6}$$

where $R_j^\star$ denotes the conjugate of R_j, and $\mathbb{T}_i(\mathbf{A})$ denotes the tangent space of low-rank manifold *w.r.t* R_i to the point $\mathbf{A}$.

Theorem (Exact-Recovery Condition)

The estimated $\hat{\mathbf{A}}_i$, obtained by Reshuffled-TD, are equal to the true $\mathbf{A}_i^*$ for all i, when

$$\max_{i=1,\ldots,N} \mu_i(\mathbf{A}_i^\star) < \frac{1}{3N-2}, \tag{7}$$

where N denotes the number of the components.

Reshuffled Tensor Decomposition

Application: Image steganography

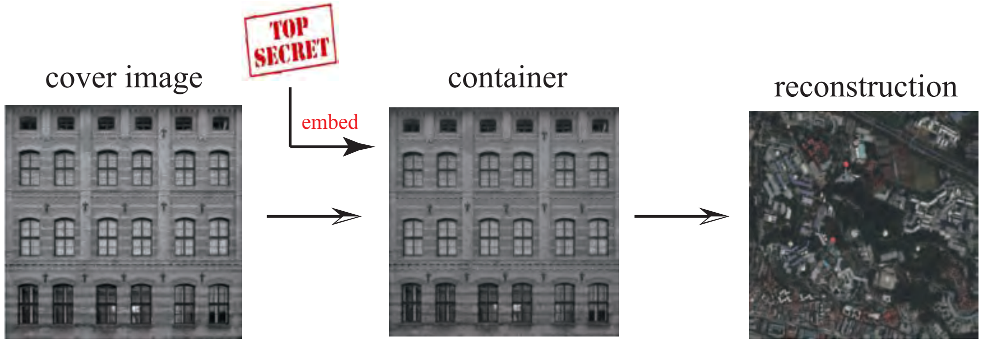
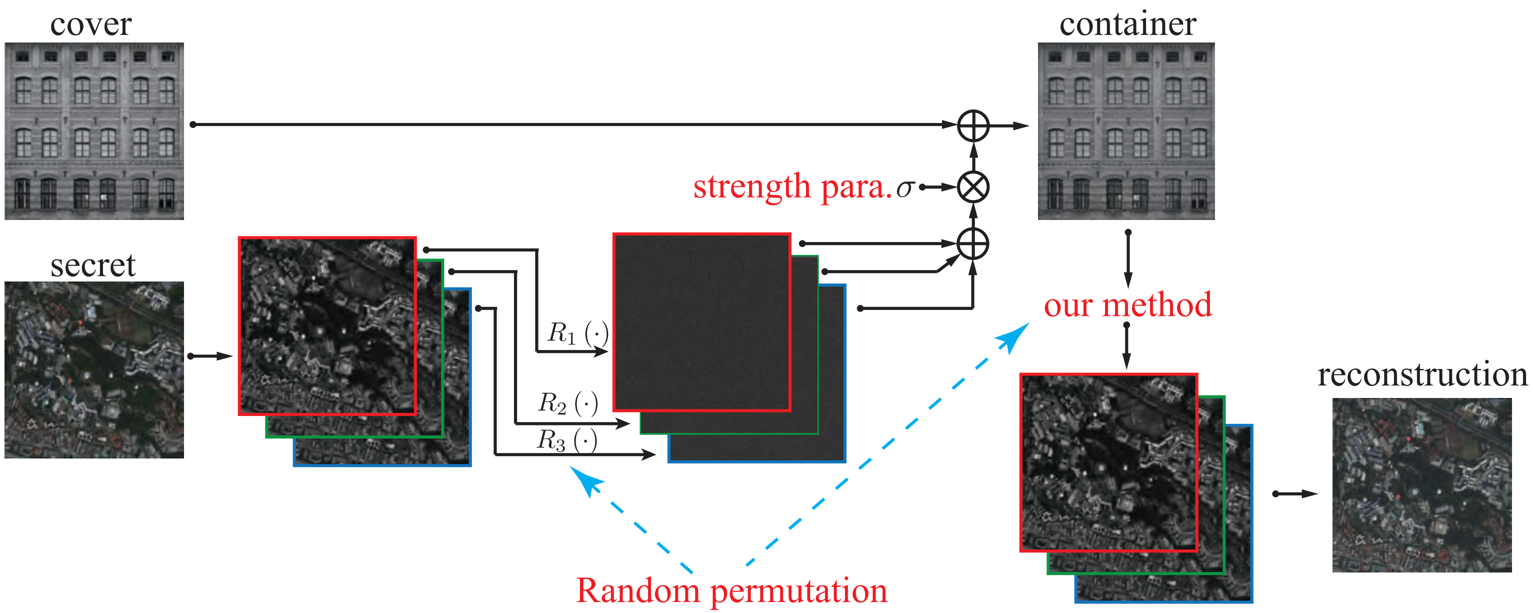


Figure Illustration of image steganography.

System design:



Reshuffled Tensor Decomposition

Result illustration:

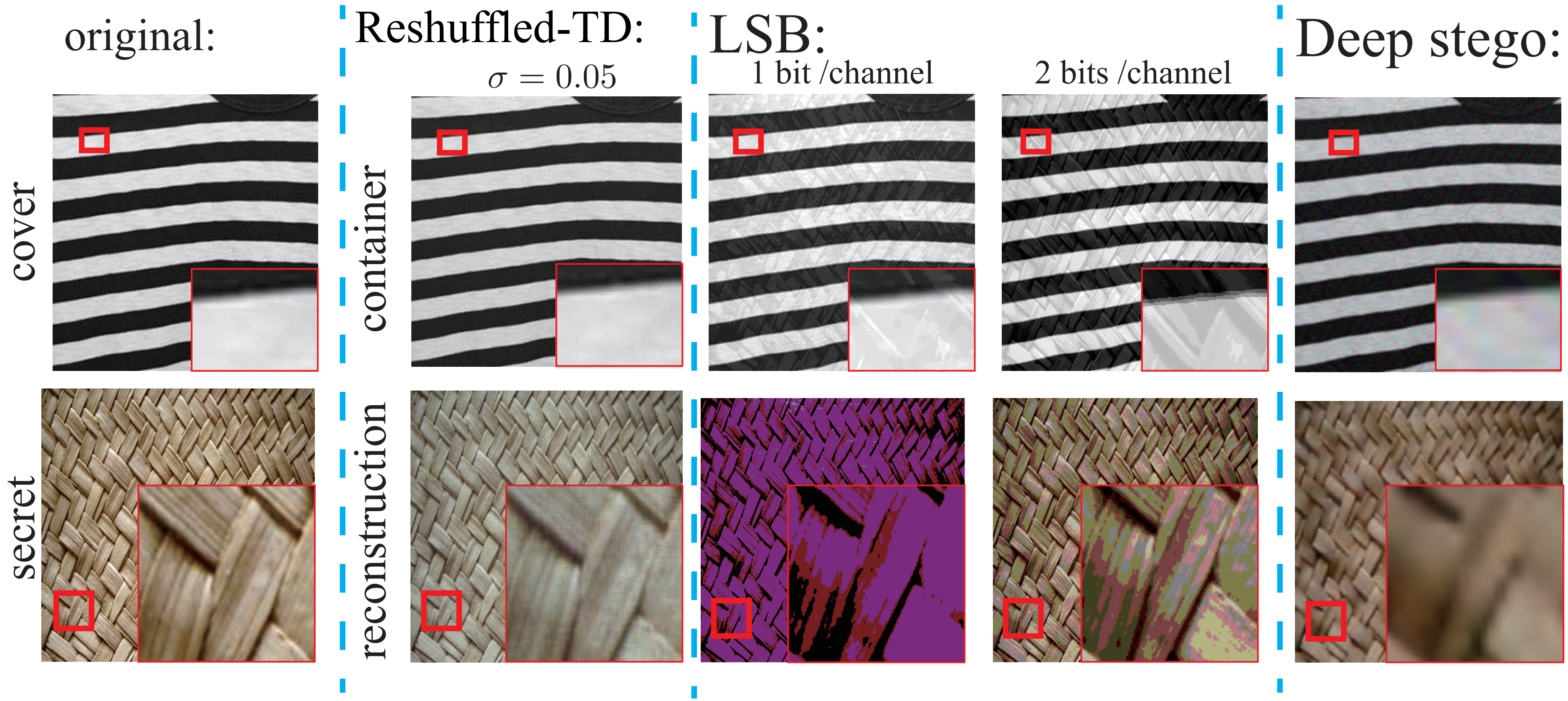


Figure Example of the experimental results by using Reshuffled-TD, LSB and deep stego.

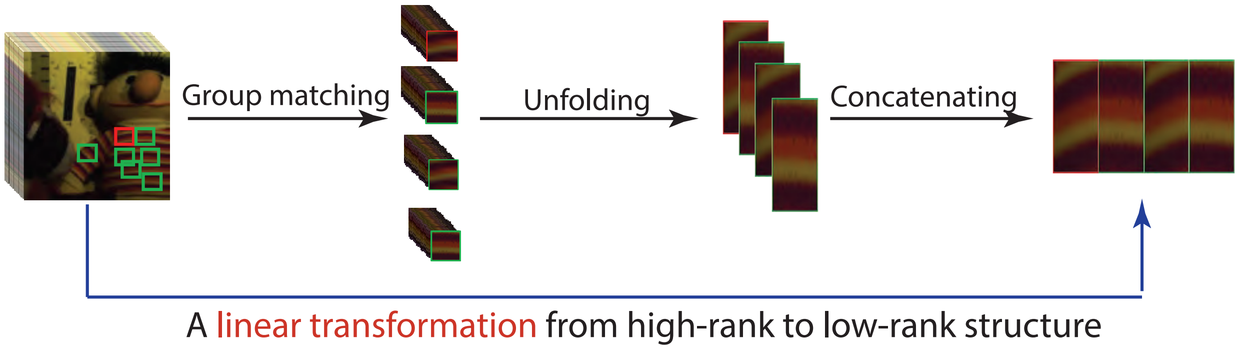
Matrix Completion under Multiple Transformation

[Li et al, CVPR'19]

Background:

- In recent computer vision tasks, the completion is usually employed on the variants of data, such as "non-local" or filtered, rather than their original forms.

An example – Non-local Trick in Image Restoration



Summary

A significant low-rank structure appears under some transformations.

Problem

The conventional theoretical analysis for guarantee is no longer suitable.

Matrix Completion under Multiple Transformation

In the simplest case, the completion problem can be solved by the following optimization problem:

$$\min_{\mathbf{X}\in\mathbb{R}^{m_1\times m_2}} \|\underbrace{\mathcal{Q}}_{\text{Linear transformation}}(\mathbf{X})\|_* \quad s.t. \|\mathcal{P}_\Omega(\mathbf{X}) - \mathcal{P}_\Omega(\mathbf{Y})\|_F \leq \delta,$$

Theorem

With some assumptions on the $\mathcal{Q}_i, i \in [K]$, and further assume that the tuning parameter satisfies $\lambda > \|P_\Omega(\eta)\|_2/\sqrt{M}$. Then the reconstruction error is upper-bounded by

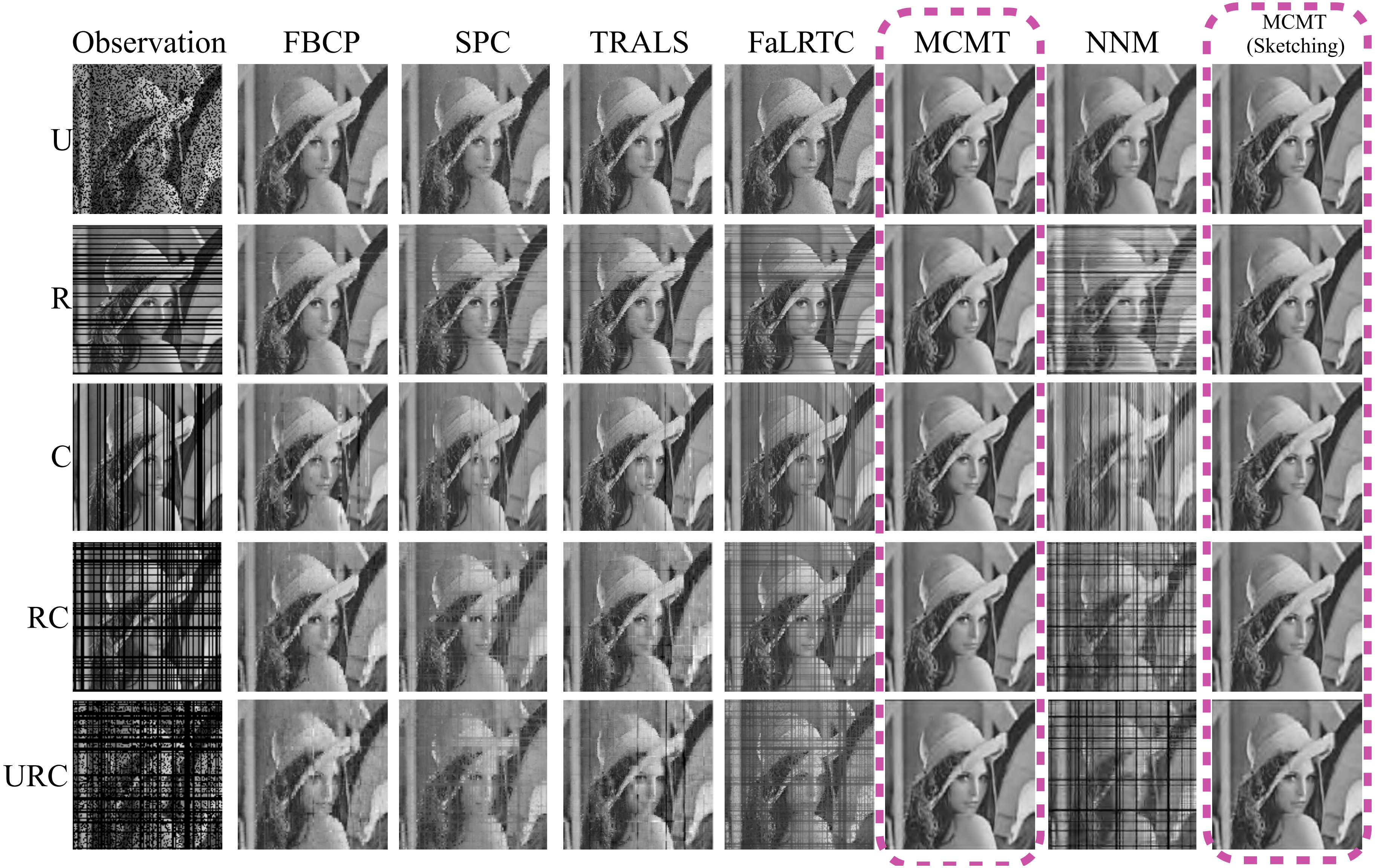
$$\|\hat{\mathbf{M}} - \mathbf{M}_0\|_F \leq \mathcal{O}\left(\lambda \cdot M^{0.5}\frac{\delta_{\max}(\{\mathcal{Q}_i\})}{\delta_{\min}(\{\mathcal{Q}_i\})}\left(K^2 + M^{K-0.5}\delta_{\max}(\{\mathcal{Q}_i\})\right)\right), \tag{2}$$

where $\delta_{\max}(\cdot)$ and $\delta_{\min}(\cdot)$ denotes the maximum and the non-zero minimum singular values from all $\mathcal{Q}_i$'s, respectively.

Remark

The upper-bound of the reconstruction error is linearly controlled by the condition number of the transformations.

Illustrative Experiment

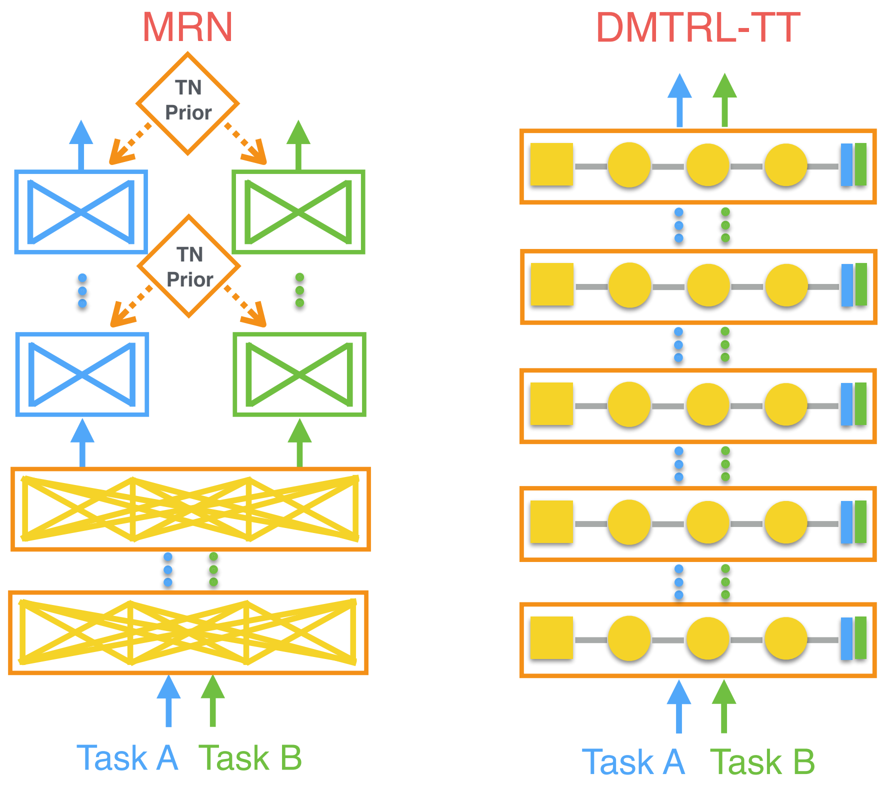


Tensor Networks for Model Representation

Deep Multi-task Learning

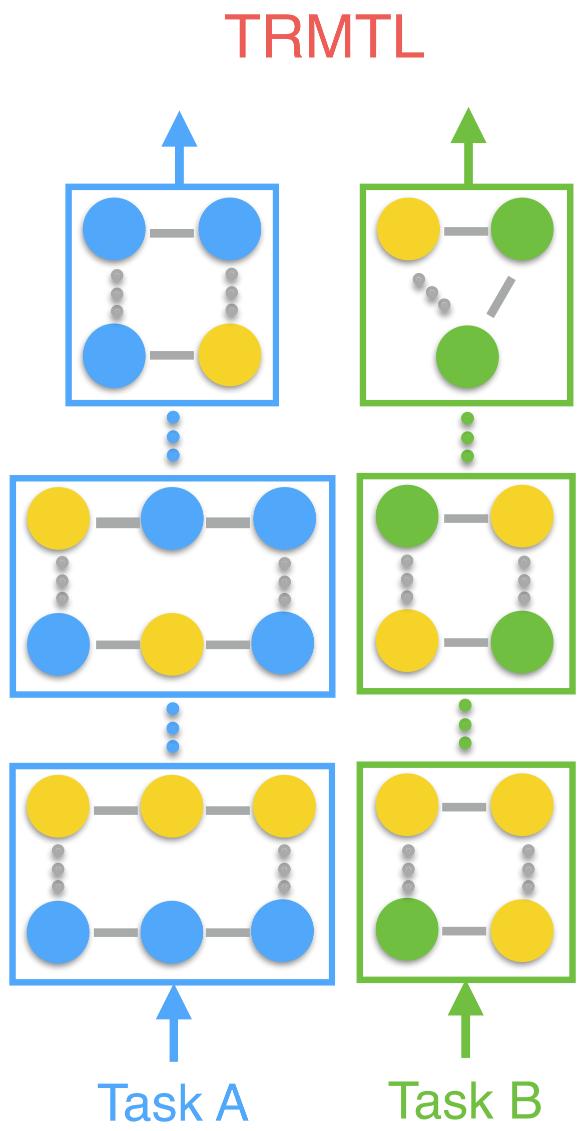
- Cannot handle data from multiple sources/modalities
- Cannot consider heterogeneous networks for individual task
- Lack flexibility in knowledge-sharing mechanism

[Long et al. NIPS 2017] *[Yang et al, ICLR 2017]*



Tensor Ring Multi-task Learning

- Heterogeneous DNN for each task
- Flexibility in knowledge-sharing pattern
- High efficiency by sharing information in latent space
- Disadvantages: choosing the number and location of cores for sharing is difficult.



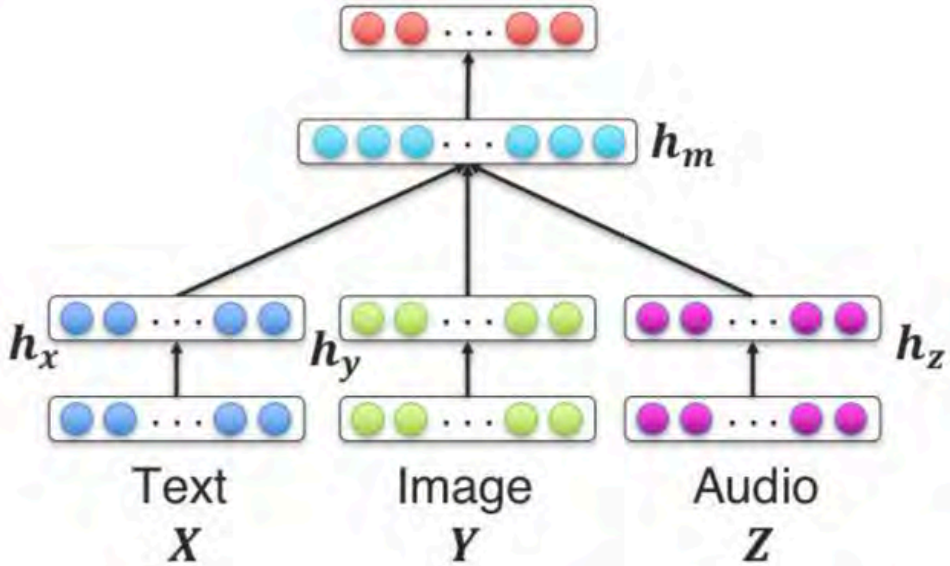
Multimodal Learning

Joint Multimodal Representation

Simply concatenates all three individual representations:

$$h_m = f(W \cdot [h_x, h_y, h_z])$$

- ➢ Similar to early fusion

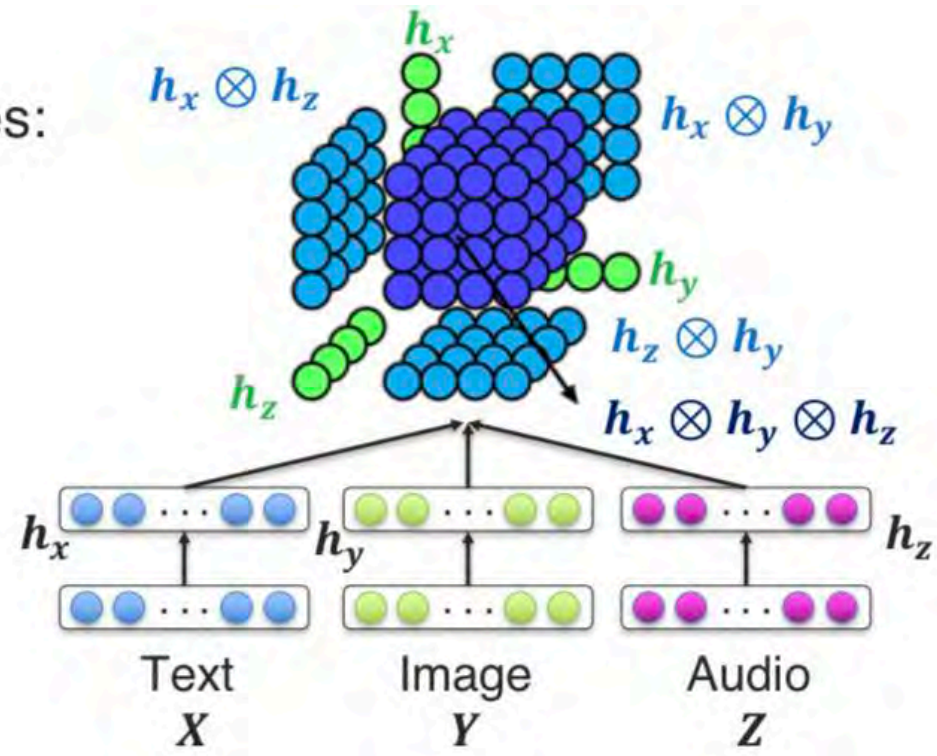


Multimodal Tensor Fusion Network

Can be extended to three modalities:

$$\boldsymbol{h}_m = \begin{bmatrix}\boldsymbol{h}_x \\ 1\end{bmatrix} \otimes \begin{bmatrix}\boldsymbol{h}_y \\ 1\end{bmatrix} \otimes \begin{bmatrix}\boldsymbol{h}_z \\ 1\end{bmatrix}$$

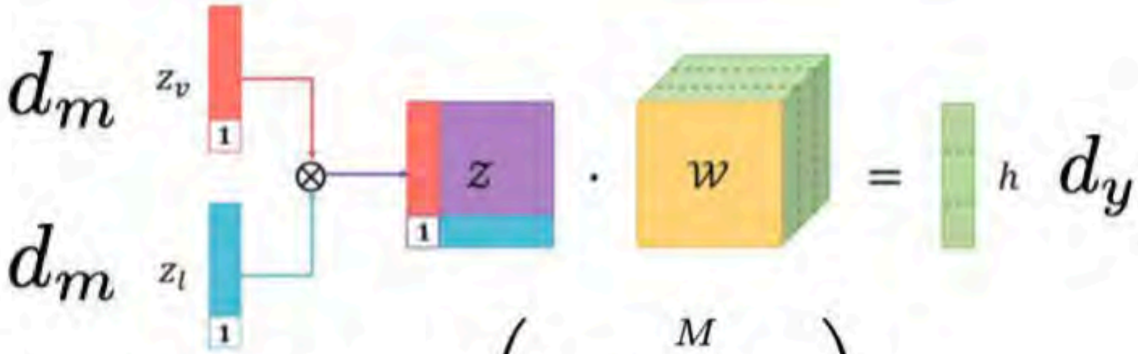
Explicitly models unimodal, bimodal and trimodal interactions!



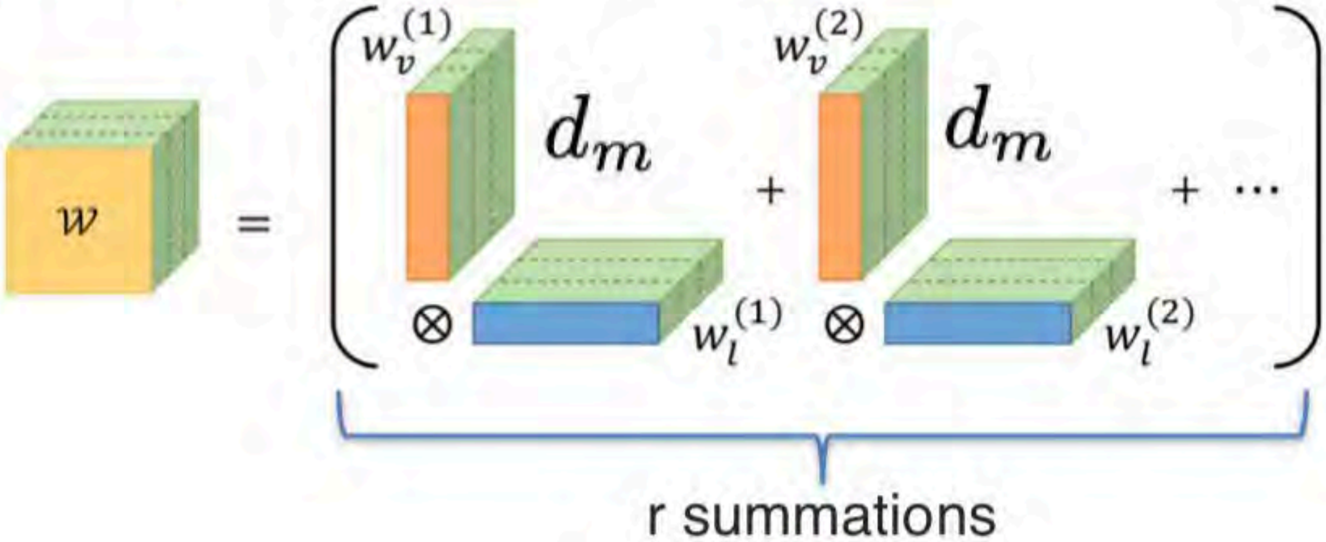
[Liu et al, ACL 2018]

Low-rank Tensor Fusion

[Liu et al, ACL 2018]

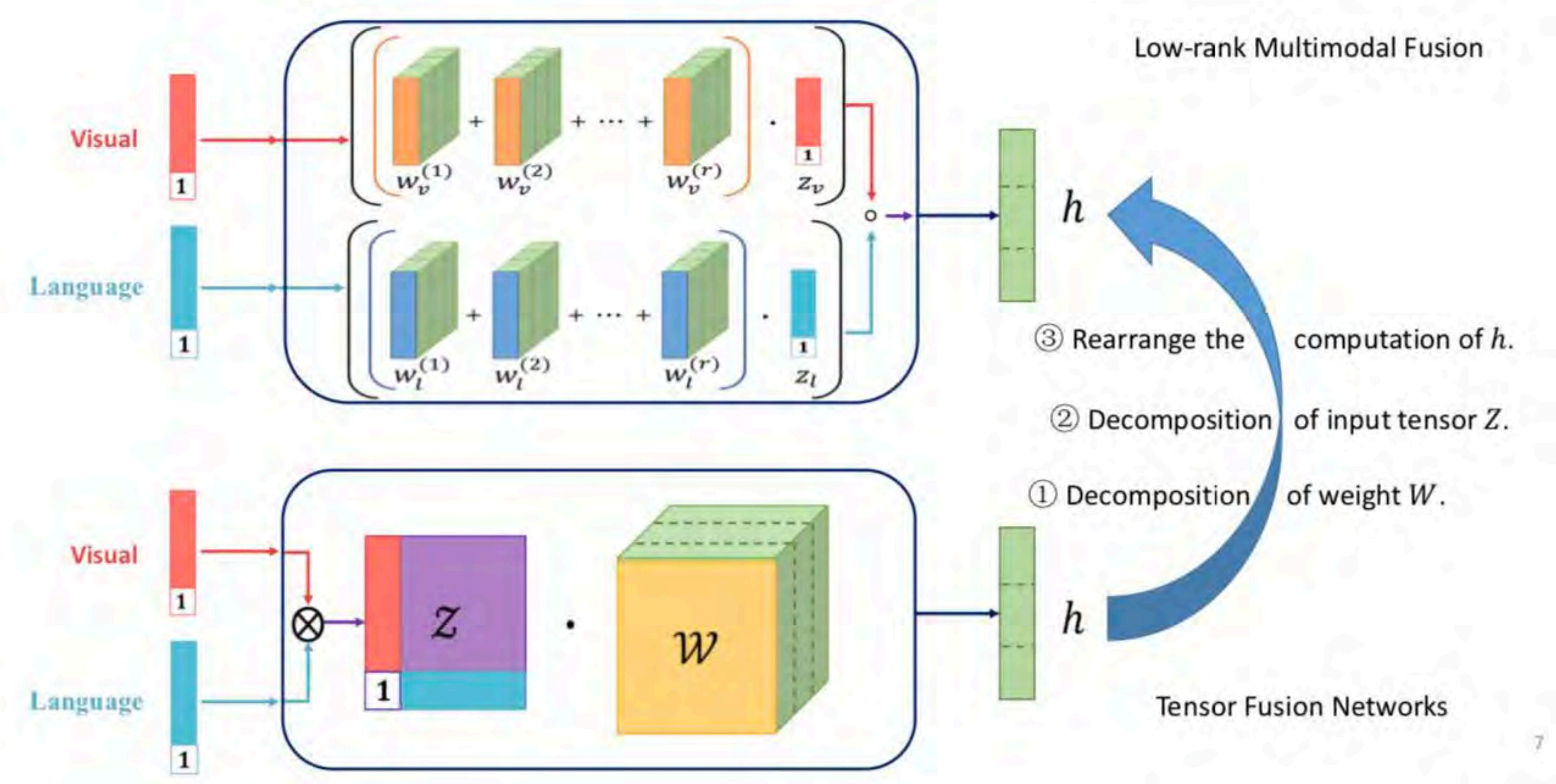


- Rank-r approximation $O\left(d_y \times \prod_{m=1}^{M} d_m\right)$



Low-rank Tensor Fusion

[Liu et al, ACL 2018]



Imperfect Time Series Data

[Liang, ACL 2019]

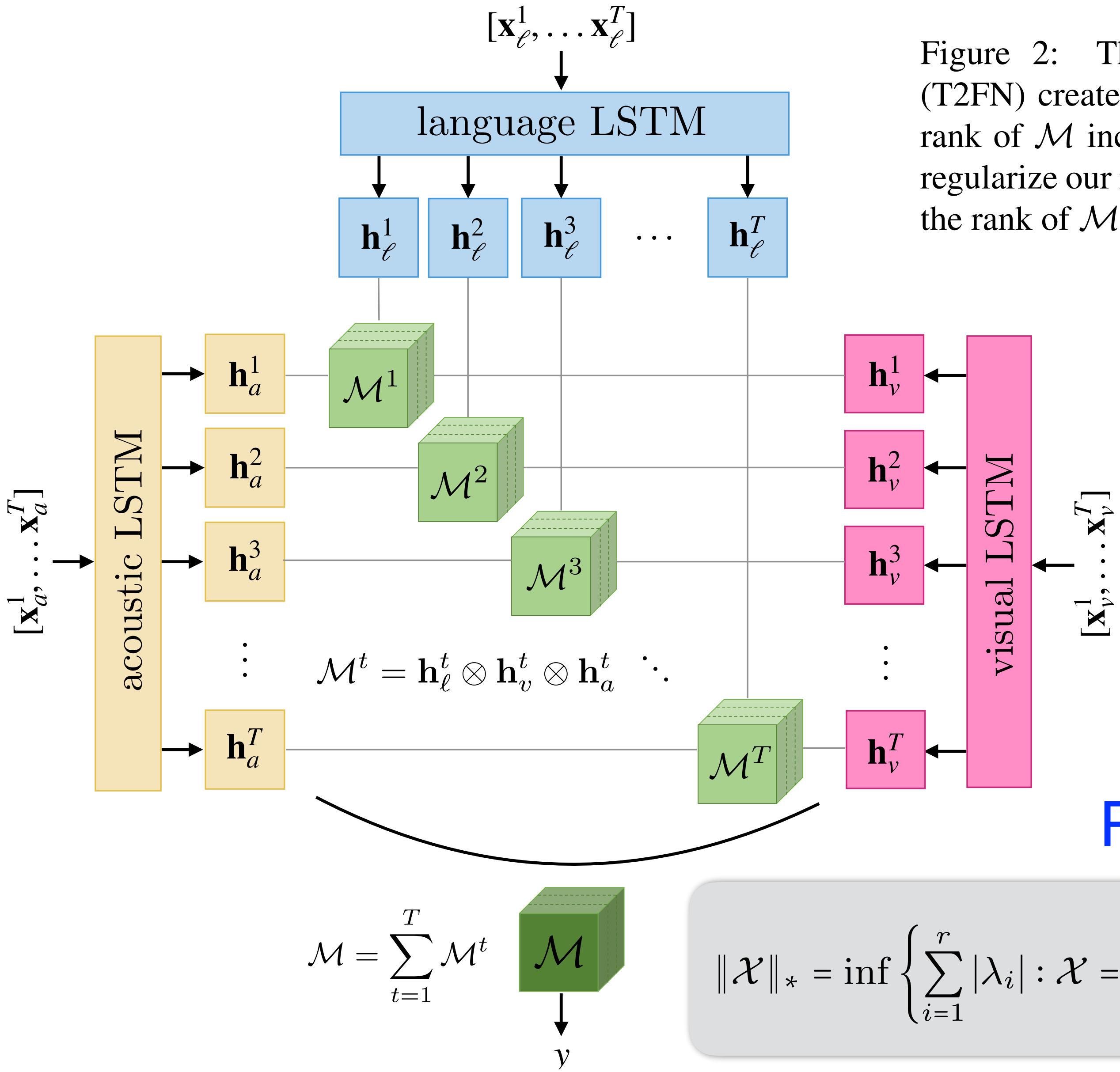


Figure 2: The Temporal Tensor Fusion Network (T2FN) creates a tensor $\mathcal{M}$ from temporal data. The rank of $\mathcal{M}$ increases with imperfection in data so we regularize our model by minimizing an upper bound on the rank of $\mathcal{M}$.

Tensor Rank Regularization

$$\|\mathcal{X}\|_* = \inf\left\{\sum_{i=1}^{r}|\lambda_i| : \mathcal{X} = \sum_{i=1}^{r}\lambda_i\left(\bigotimes_{m=1}^{M}\mathbf{w}_m^i\right), \|\mathbf{w}_m^i\| = 1, r \in \mathbb{N}\right\}$$

Open problems

- Tensor network expressive power analysis
- Learning of tensor network structure
- Fast algorithms for tensor network representation

- What challenging problems in machine learning can be solved by tensor network?