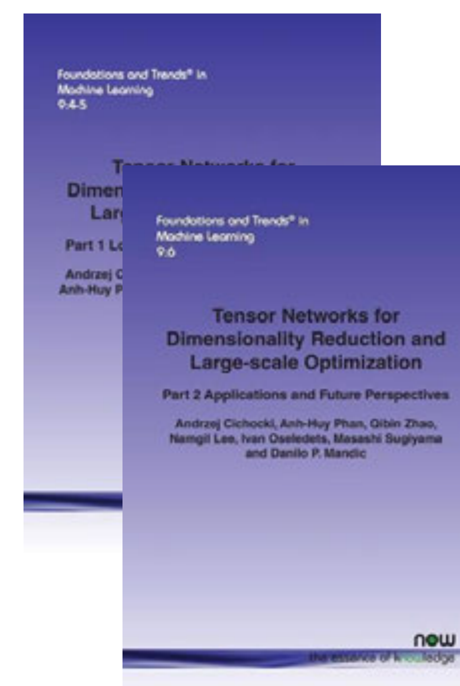
Tensor Network Representation for Machine Learning - Recent Advances and Perspectives

Qibin ZHAO

Tensor Learning Unit
RIKEN AIP

AIP Symposium
(Mar. 19, 2019)



Tensor Learning Unit - Members

Postdoctoral Researchers (2)

- Ming Hou, Chao Li

Part-timer (2)

- Longhao Yuan (PhD student), Xuyang Zhao (PhD student)

Interns (4)

- Canada, Japan, China

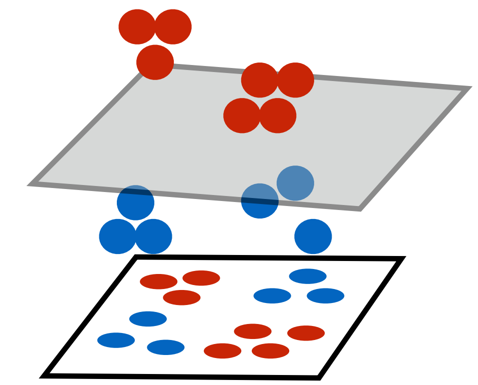
Visitors (9)

- Andrzej Cichocki, Toshihisa Tanaka, Jianting Cao
- Guillaume Rabusseau, Justin Dauwels, Danilo Mandic, Brahim Chibdraa, Cesar F. Caiafa, Jordi Sole Casals

Background and Problems

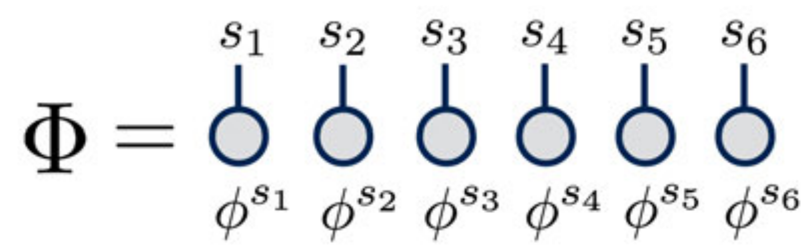
Kernel learning

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



Kernel Learning

- Problems become easier when mapping to higher dimensional space.
- Curse of dimensionality, grows exponentially
- Weights can be exponentially big
- "kernelization" scales quadratically with training set size. In the era of big data, this issue is cited as one reason why neural nets have overtaken kernel methods.
- Low generalization due to representer theorem



Rank-1 tensor

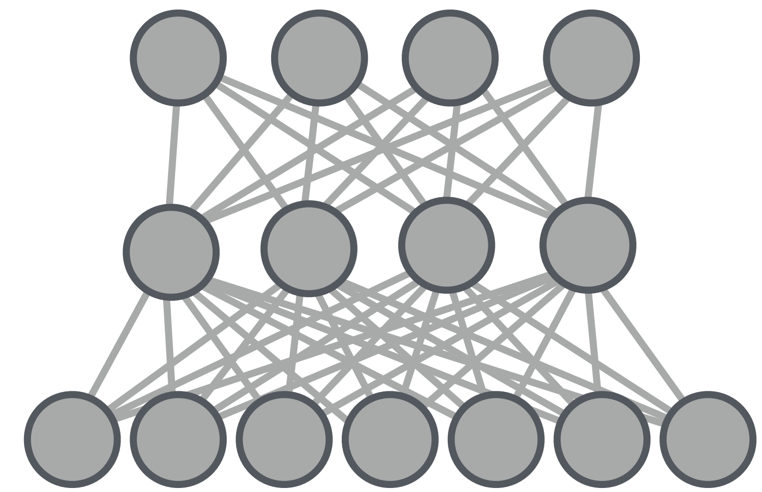
$$W = \sum_j \alpha_j \Phi(x_j)$$

Perfect Problem for Tensor Networks to solve

Background and Problems

Neural Networks

- Weight matrix is huge but highly redundant.
- Low-rank compression: limited compression rate
- Computational inefficient due to huge parameters
- Not applicable for small devices



Neural Nets

Multi-modal deep learning, multi-task deep learning

Tensor Networks is a natural tool to solve these problems

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

Neural Network (NN) vs. Tensor Network (TN)

Similarity

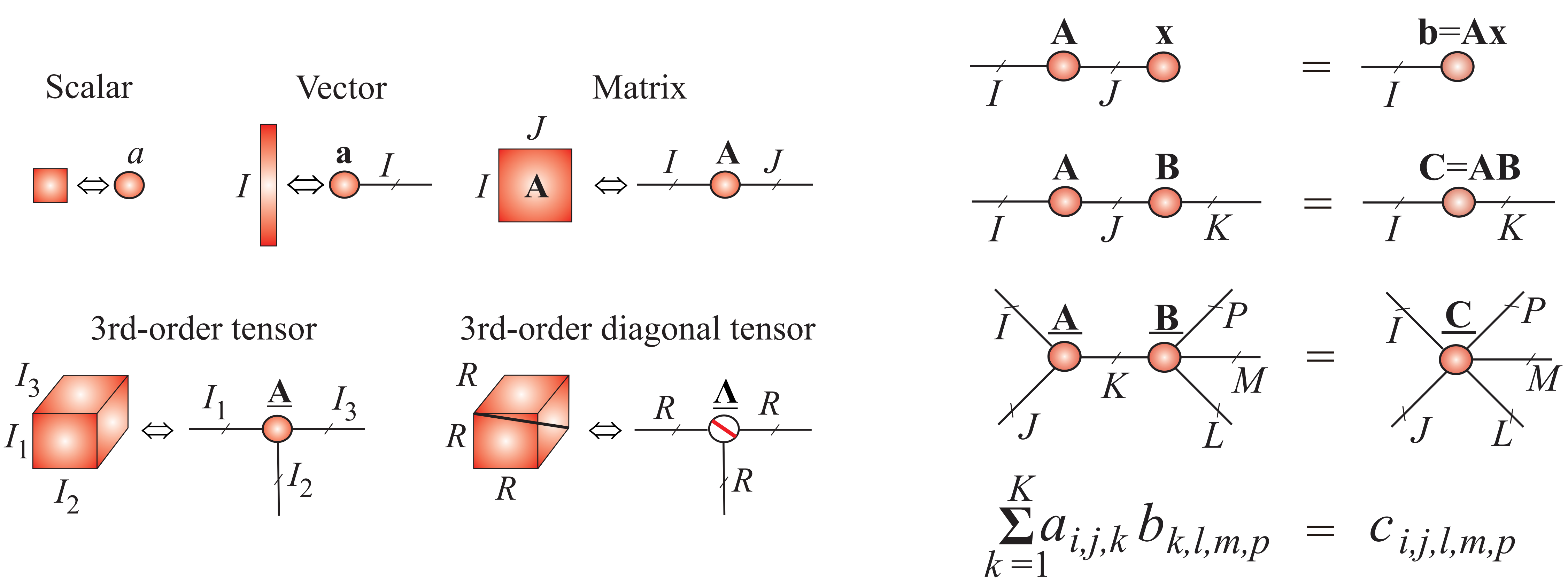
- Assembling simple units (neurons or tensors) into complicated functions

Difference

- Decision functions in ML vs. wavefunctions in quantum mechanics
- Nonlinear in NN vs. linear in TN
- NN do non-linear things to low-dimensional space vs. TN do linear things in high-dimensional space

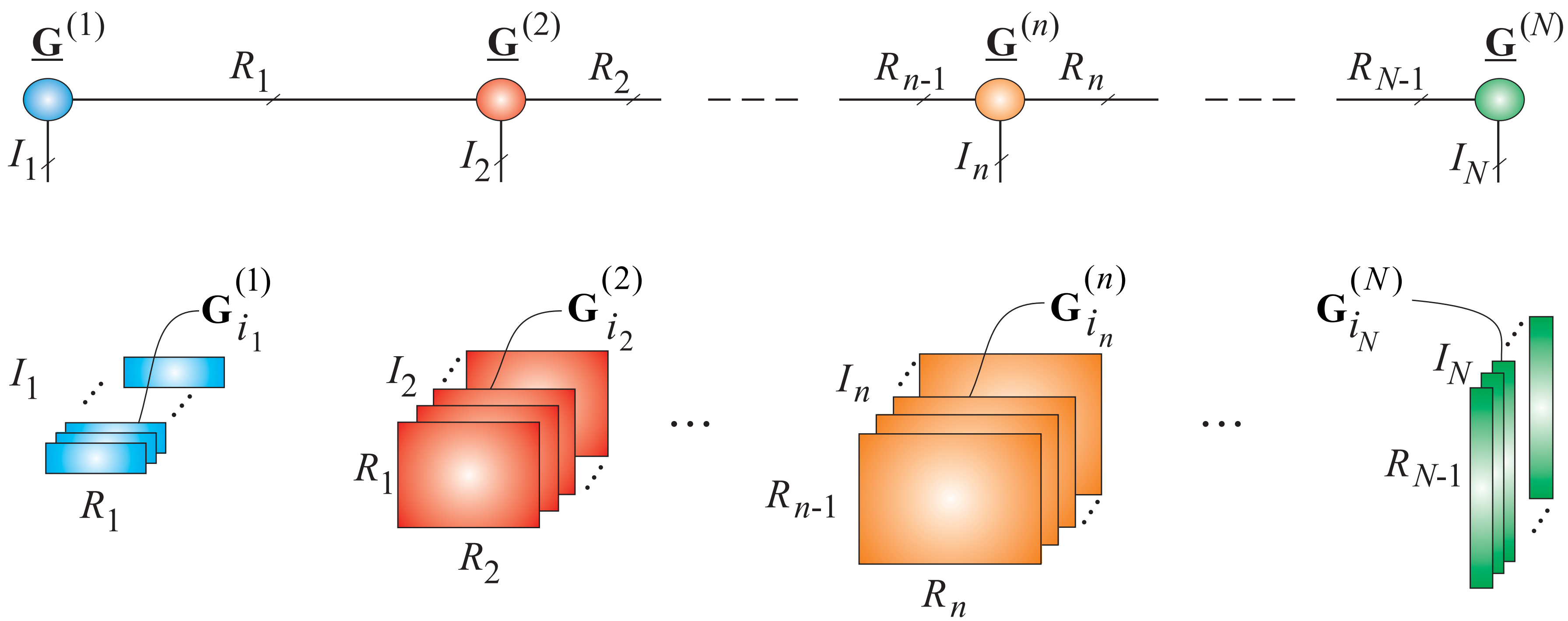
What Are Tensor Networks (TNs) ?

- 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]

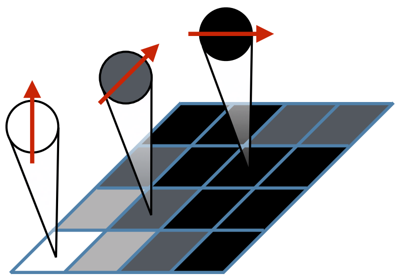


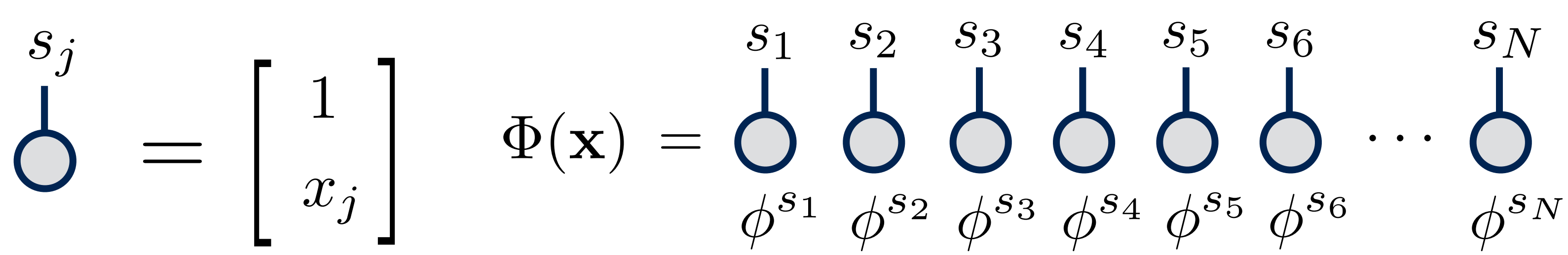
TT: tensor train decomposition; MPS: matrix product state

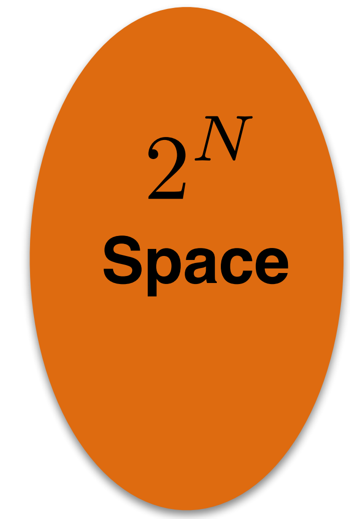
- Efficient to represent I^N data values by $\mathcal{O}(NIR^2)$ parameters
- Efficient to compute or optimize TT/MPS by DMRG algorithm

TNs for Weight Compression & Kernel Learning

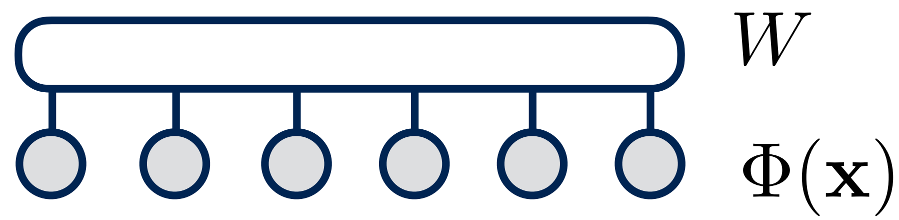
- Input: $\mathbf{x} = [x_1, \; x_2, \; x_3, \; \ldots \; , \; x_N]$ *[E. Stoudenmire, NIPS 2016]*
- Nonlinear mapping by tensor product (Hilbert space)





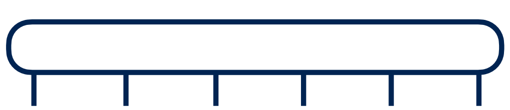


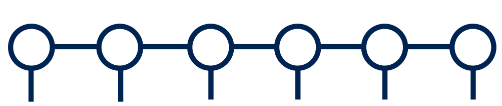
- Decision function - W is an *Nth*-order tensor

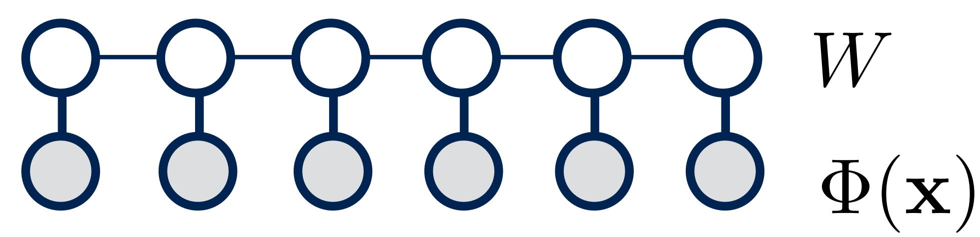
$f(\mathbf{x}) = W \cdot \Phi(\mathbf{x}) =$ 

- TT representation of weight parameter

[A. Novikov, NIPS 2015]

$W =$ 

$\approx$ 

$f(\mathbf{x}) =$ 

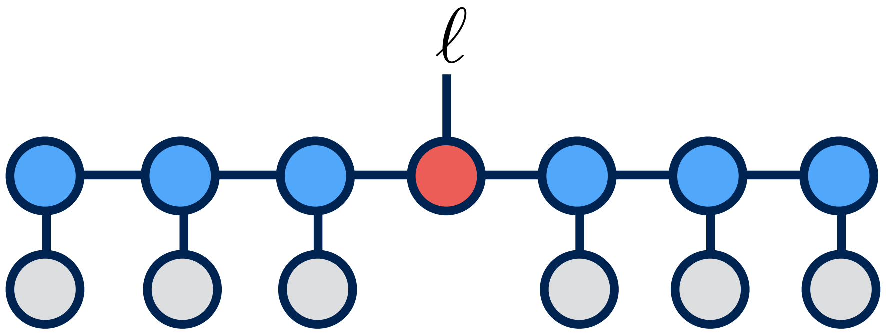
TNs for Weight Compression & Kernel Learning

- Optimization algorithm scaling: $O(NN_T m^3)$ *[E. Stoudenmire, NIPS 2016]*

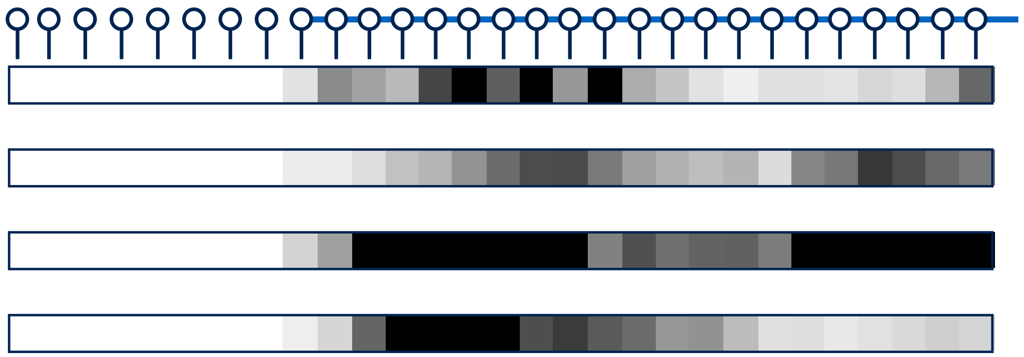
  m: TT rank, N_T : Sample size

- Without "kernel trick", avoiding N_T^2 scaling problem
- Without deep layers transformation
- Feature sharing for multi-class function

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

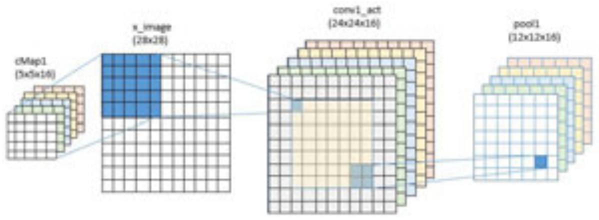
$f^{\ell}(\mathbf{x}) =$ 

- Adaptive learning

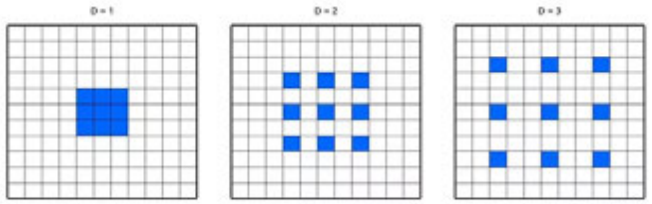


Theoretical Analysis of ConvNets

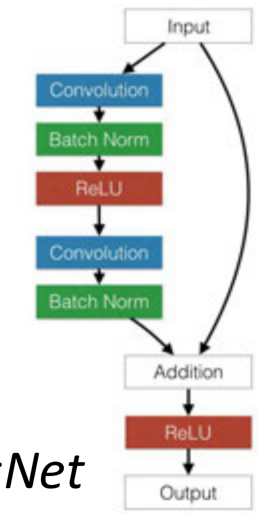
Fundamental theoretical questions:



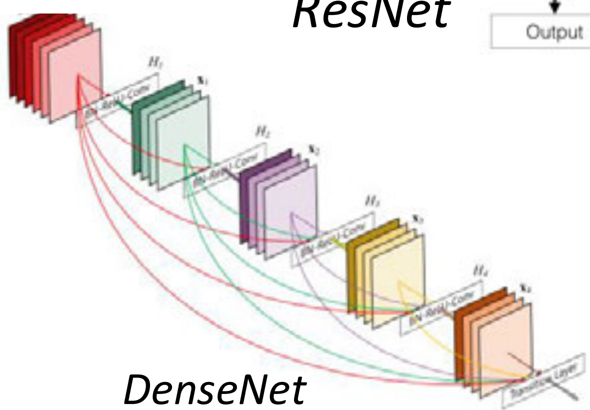
- Are deep networks efficient w.r.t. shallow one for ConvNets?
- What kind of func can different network arch represent?



- What is the inductive bias of conv/pool window geometry?
- Do overlapping operations introduce efficiency?
- Can connectivity scheme be justified in terms of efficiency?



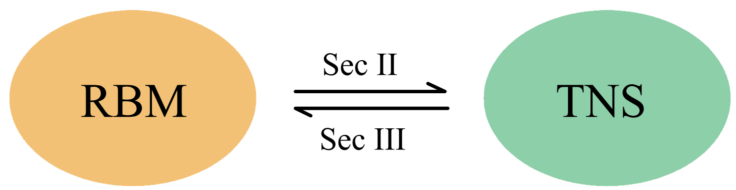
ResNet



DenseNet

Relations Between TNs and DNNs

- Equivalence of Restricted Boltzmann Machines and Tensor Networks

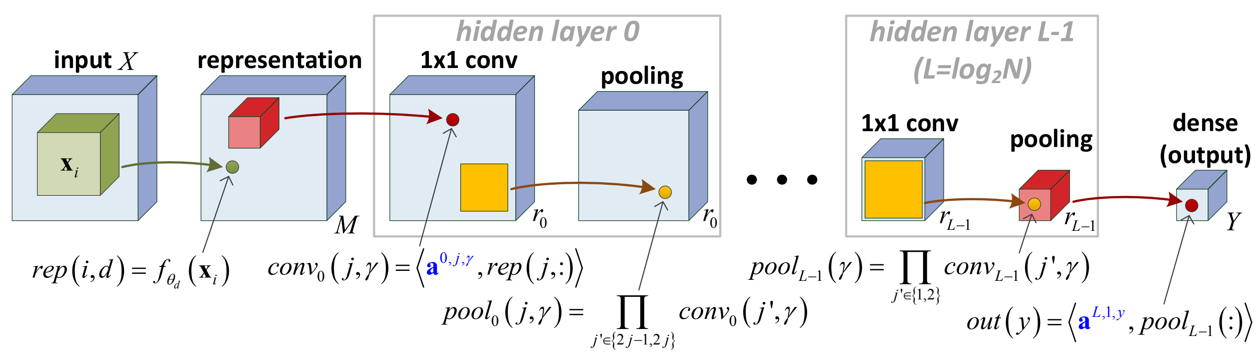


[Chen et al, Physical Review B, 2018]

[Carleo et al, Science, 2017]

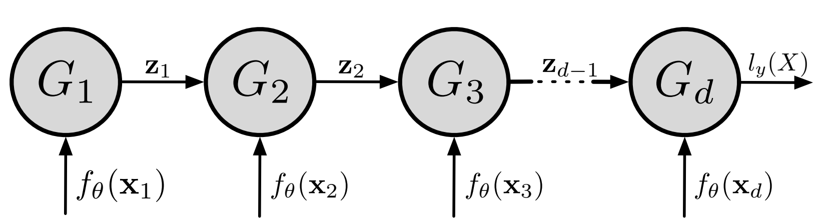
- Equivalence of Deep Convolutional Network and Hierarchical Tucker

[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]*

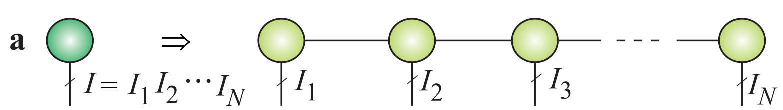


Tensor Decompositions	**Deep Learning**
CP-decomposition	shallow network
TT-decomposition	RNN
HT-decomposition	CNN
rank of the decomposition	width of the network

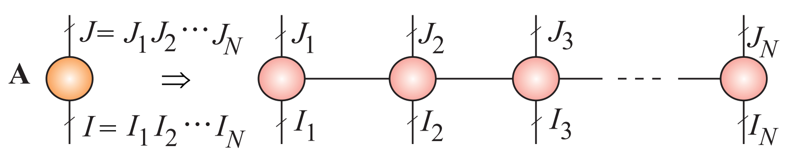
- Powerful tools to study theory behind DNN

Tensor Networks for Large-Scale Optimization Problems

- TT format of a large vector

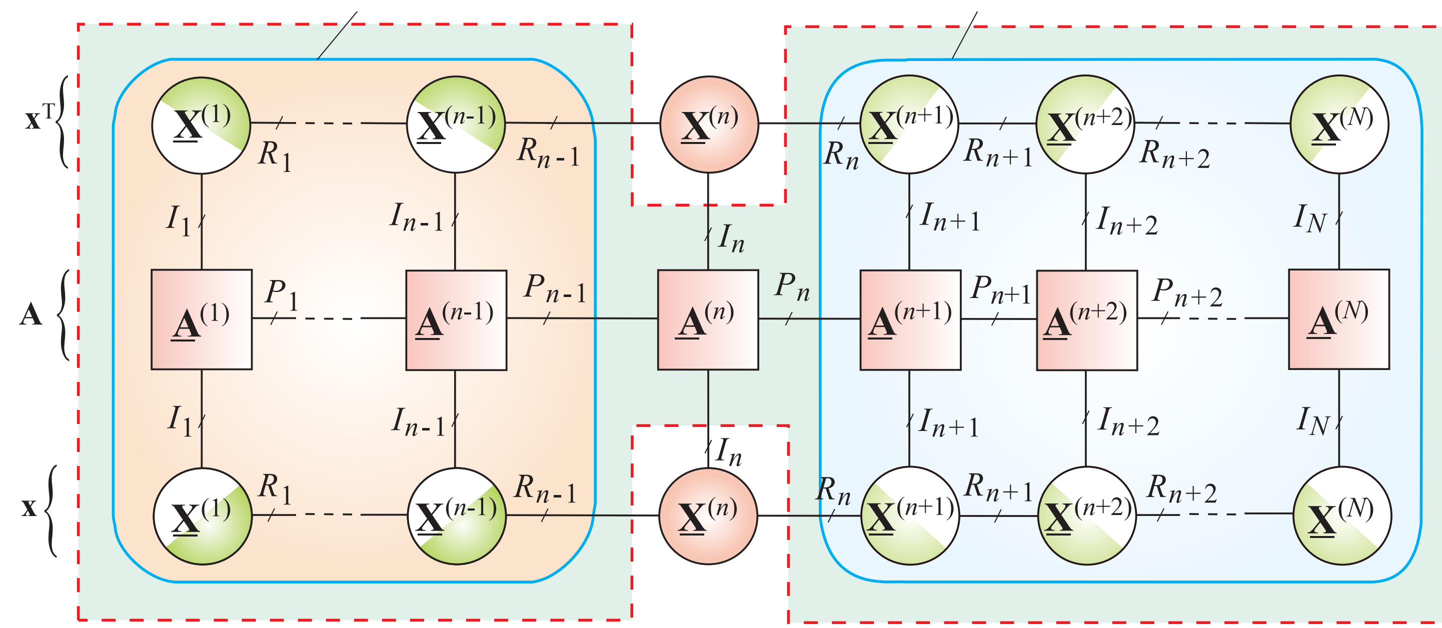


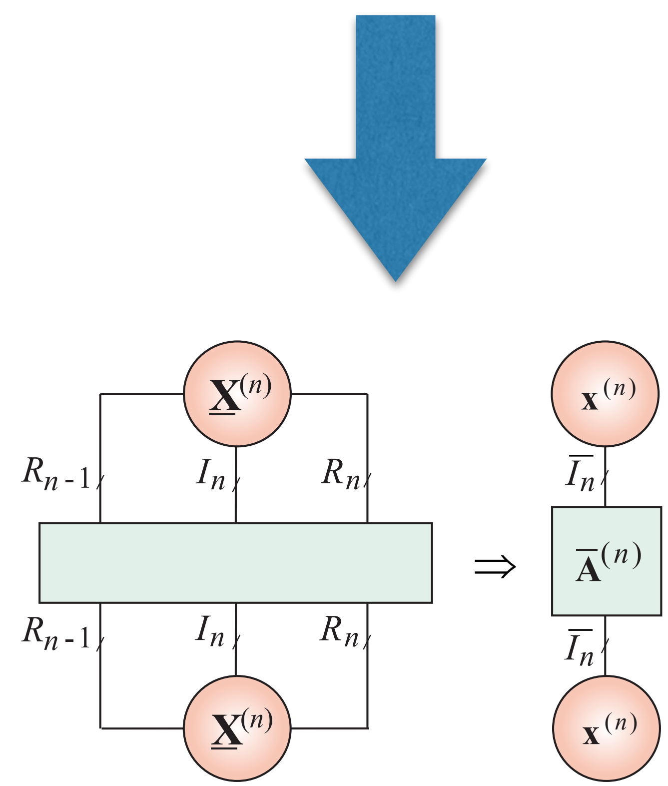
- TT format of a large matrix



- Fast ALS/DMRG algorithm
- Applicable to large-scale SVD/PCA/CCA and etc

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





Research Scheme

- Study the fundamental principle of tensor networks
- Investigate tensor networks for data representation
- Investigate tensor networks for model representation
- Explore the potential applications of tensor methods

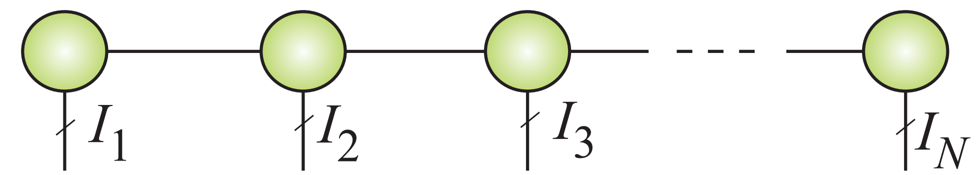
Fundamental Tensor Network Model

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

TT representation

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

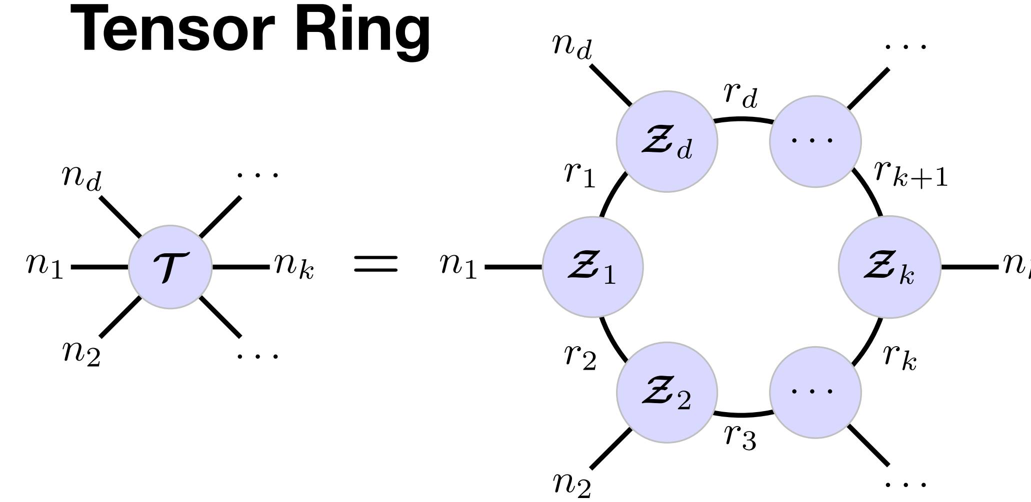
Tensor Train



Tensor ring representation

- Generalized TT without constraints on boundary cores
- Sum of TT with shared core tensors
- Efficient computation for multilinear operations
- Highly expressive model

Tensor Ring



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

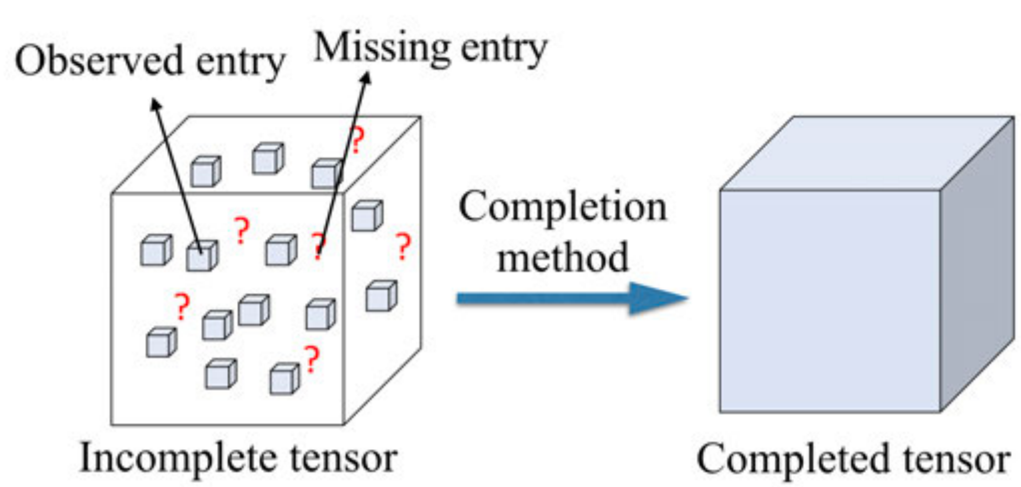
Tensor Networks for Data Representation

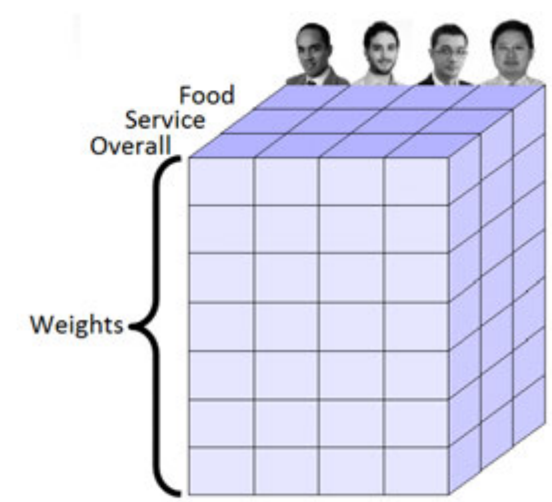
Real data is often high-dimensional

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

Real data is often incomplete

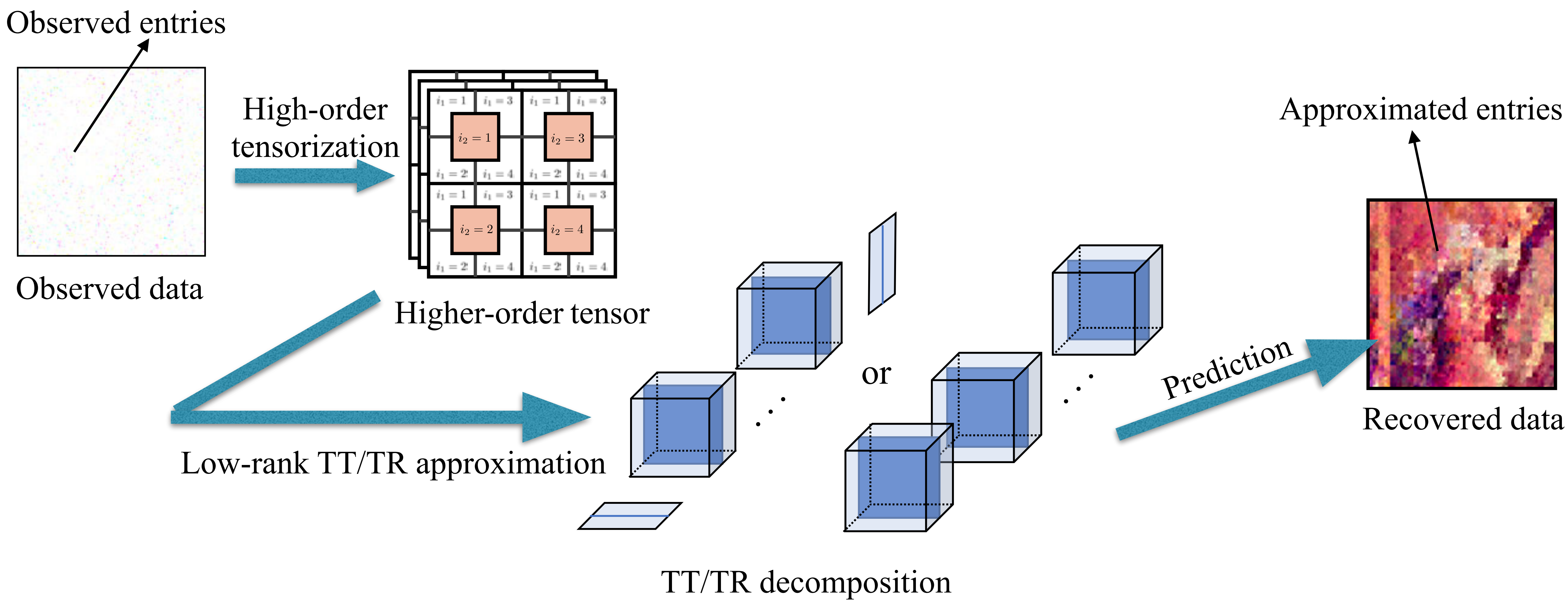
- Low-rank approximation via convex optimization (high computation cost)
- Decomposition based approach (model selection problem)
- How much structure information can be used?





Tensor Networks for Data Imputation

Tensor completion based on TT/TR decomposition



From Tensorization to Linear Transformation

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

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

[Chao et al, CVPR'19]

With mild conditions, the solution of the above problem obeys

$$\|\hat{\mathbf{M}} - \mathbf{M}_0\|_F$$
$$\leq 2\delta \cdot \frac{cond(\mathcal{Q})}{1 - \|\mathbf{R}_{\mathbf{\Lambda}}\|_2}\sqrt{\frac{\min\{n_1, n_2\}(p + \|\, [\mathcal{Q}]_{\langle 2\rangle} \,\|_2^2)}{p}}.$$

$\hat{\mathbf{M}}$ — Estimation

$\mathbf{M}_0$ — Ground truth

$cond(\cdot)$ — Condition number

$\mathbf{R}_{\mathbf{\Lambda}}$ — A matrix related to dual certificate

Beyond Unfolding: Reshuffling Operation

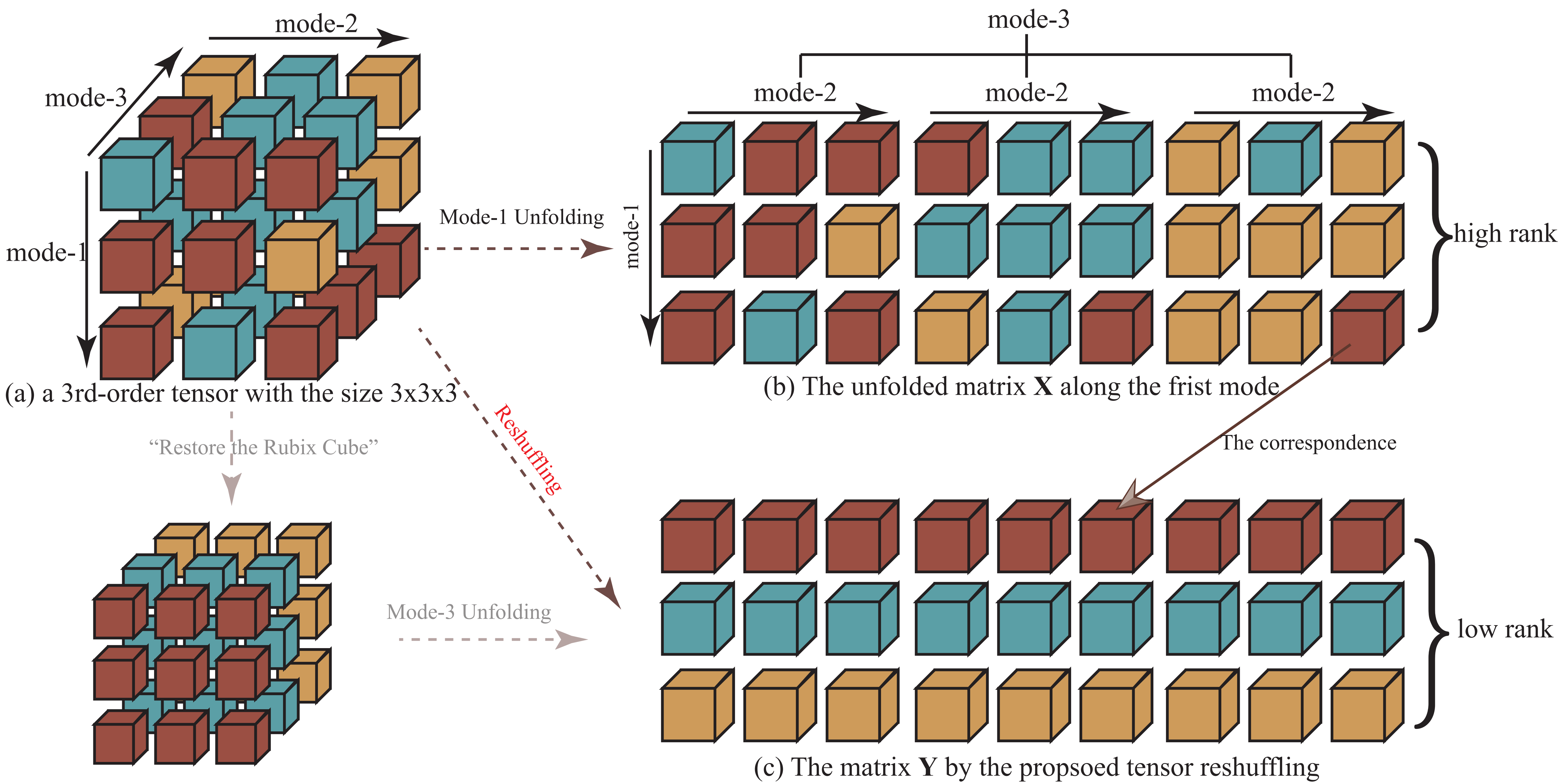
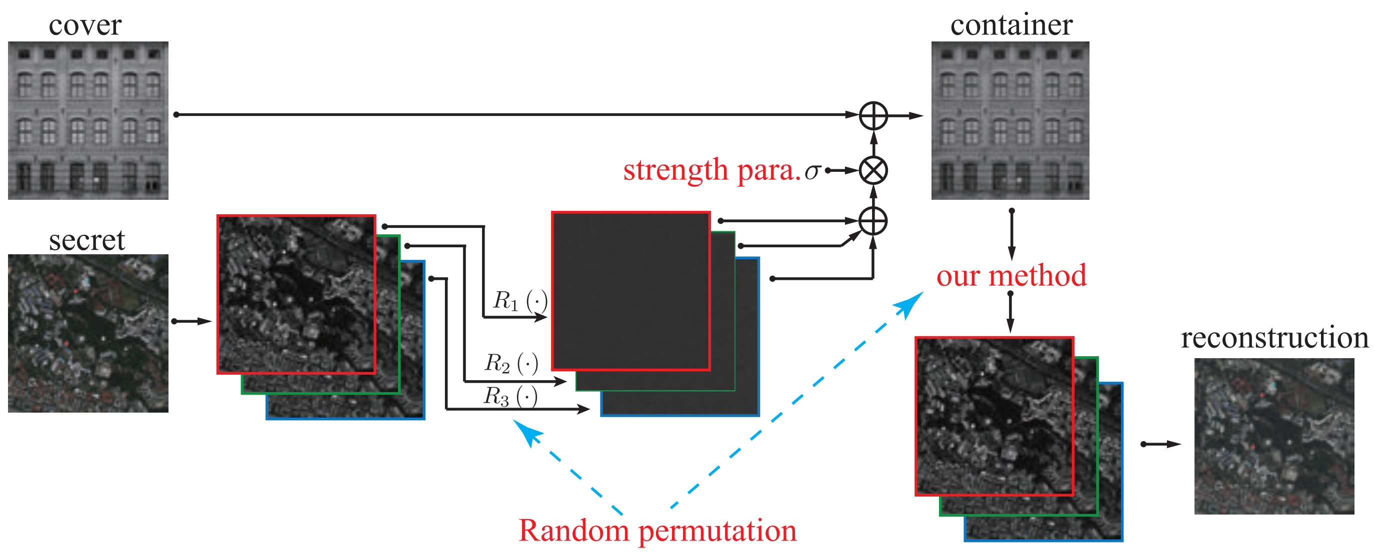


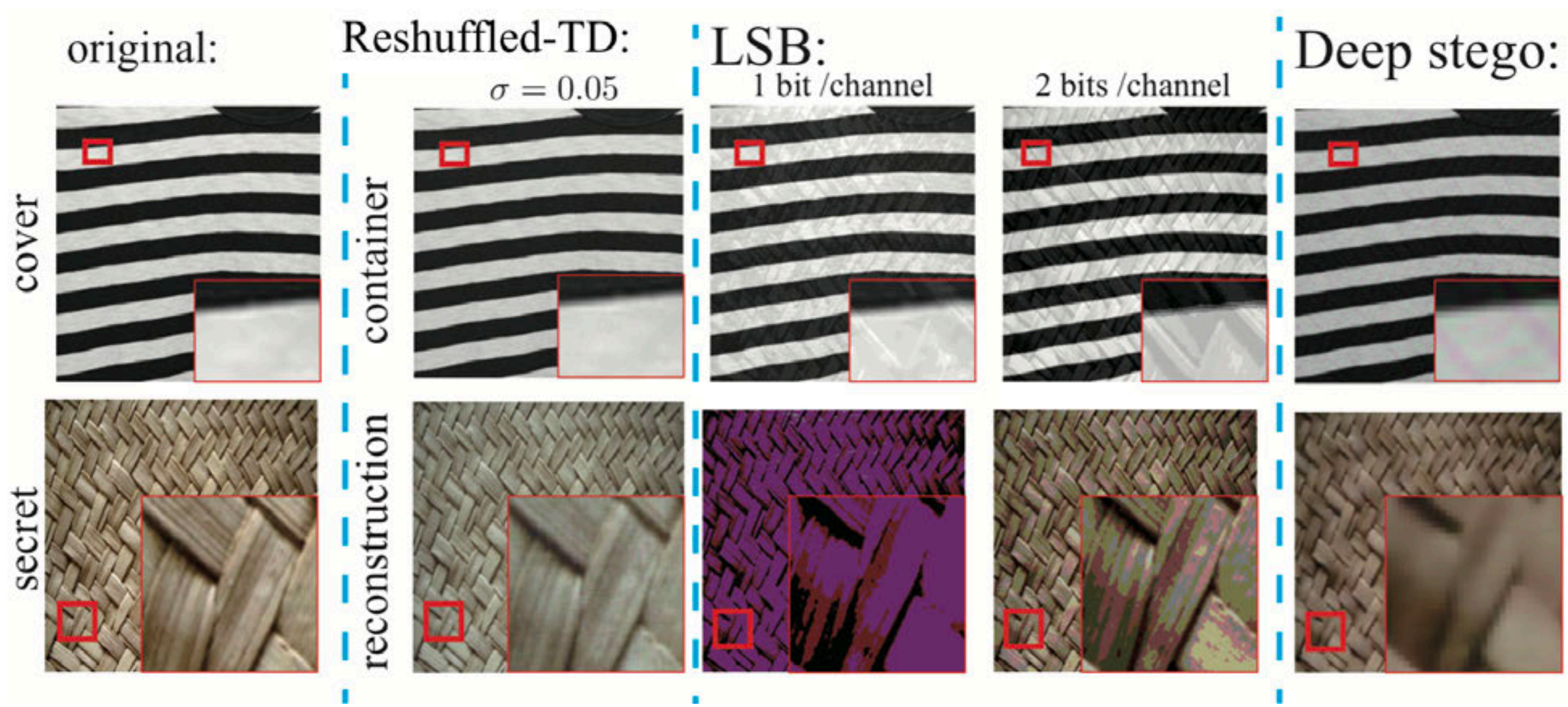
Fig. Difference between tensor unfolding and reshuffling.

Reshuffled Tensor Decomposition

Image steganography is to hide a secret image into cover image



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

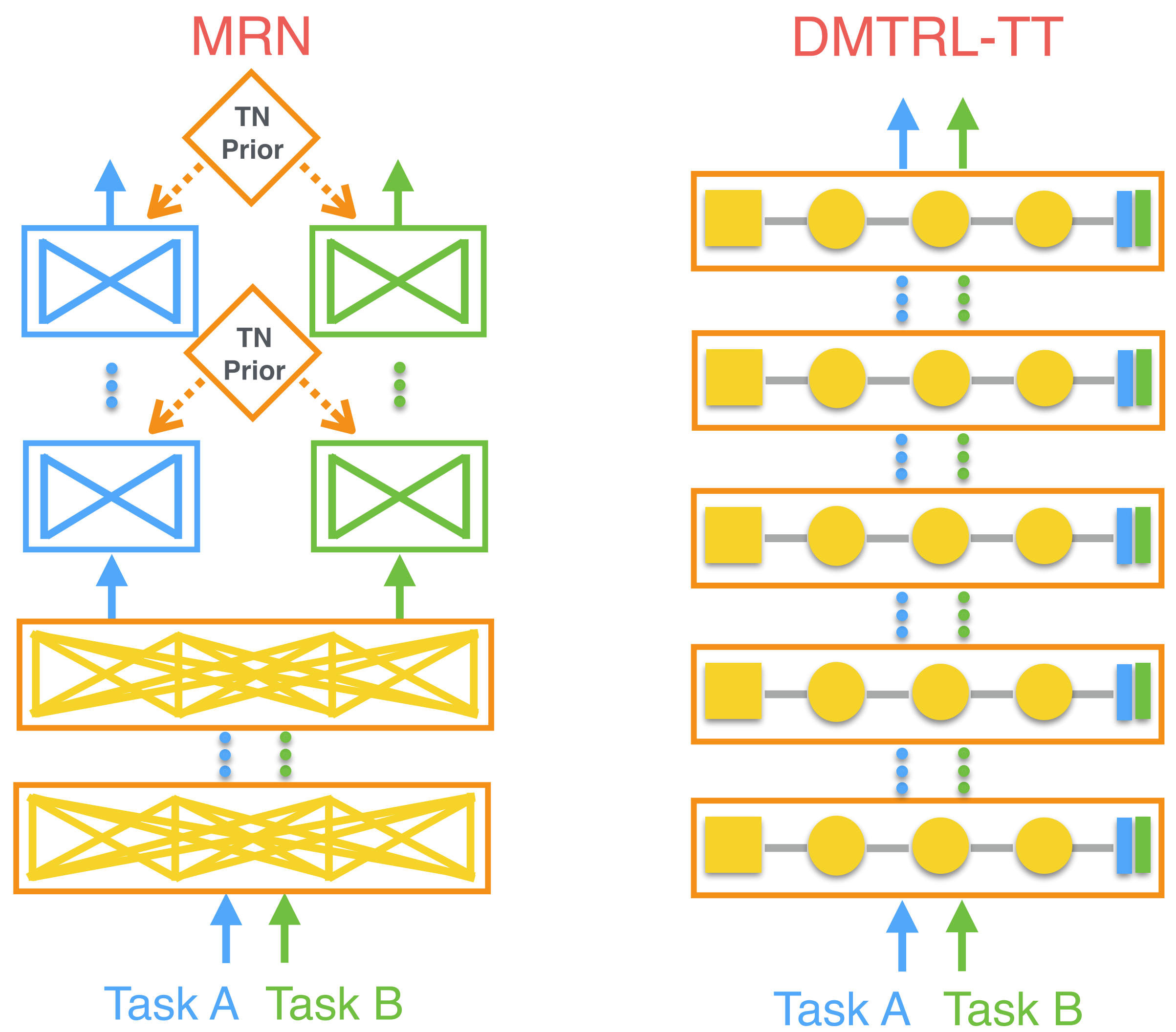


Tensor Networks for Model Representation

Deep Multi-task Learning

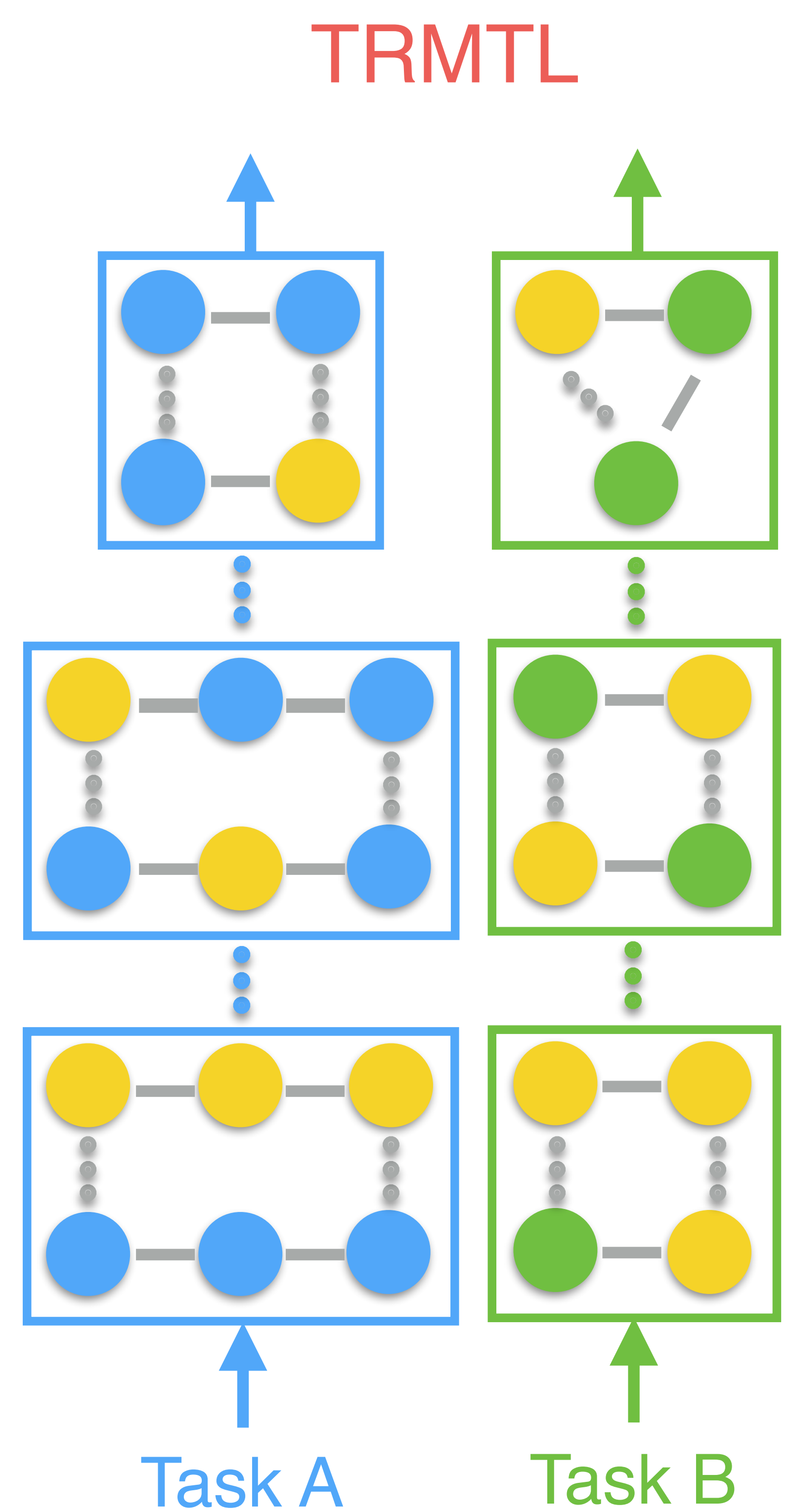
- Cannot handle data from multiple sources/modalities
- Cannot deal with heterogeneous network for individual task
- Lack flexibility in knowledge-sharing mechanism

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



Tensor Ring Multi-task Learning

- Heterogeneous DNN for each task
- Subset of TR-cores are shared among tasks
- Flexibility in knowledge-sharing pattern
- High efficiency by sharing information in latent space
- Disadvantages: choosing the best cores for sharing is difficult.



AI Support for Epileptic Diagnosis

Brain signal (EEG) examination (surface) → Drug therapy to control seizures → Brain signal (iEEG) examination (inside) → Surgical removal

Brain signal (EEG) examination (surface)	Drug therapy to control seizures	Brain signal (iEEG) examination (inside)	Surgical removal
Need to record data for one week	Not effective for all patients	Need Craniotomy and long time record	Difficult to determine the location of remove

Challenging problems

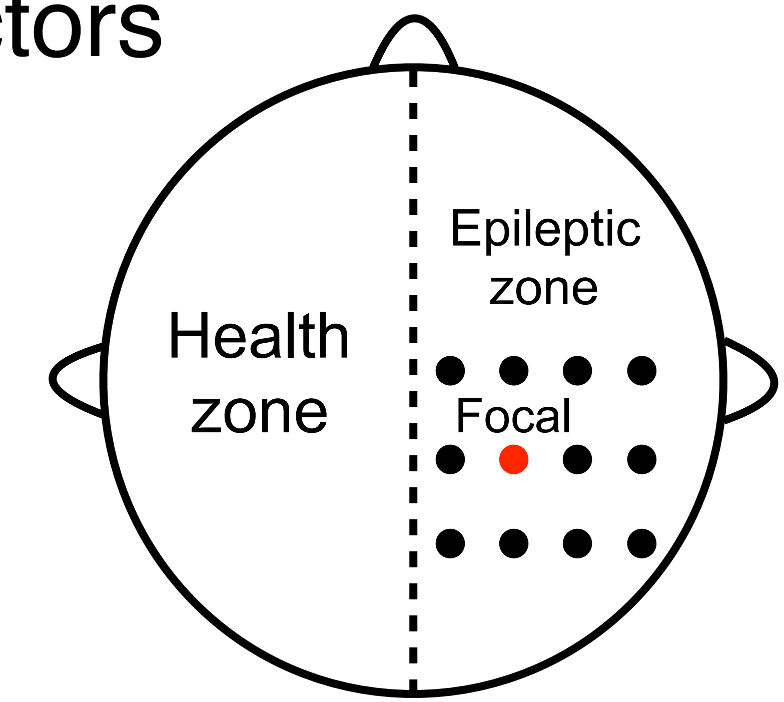
- Need special doctors, only about 600 eligible doctors in Japan.
- Need several weeks high-quality iEEG data.
- Time-consuming by several doctors’ visual judgment.
- Focal detection is not reliable.

EPILEPSY AI
JST CREST

順天堂

AI Support for Epileptic Diagnosis

- Mission: Automatic localization of epileptic focal from iEEG signals as a support technology for doctors



- High accuracy

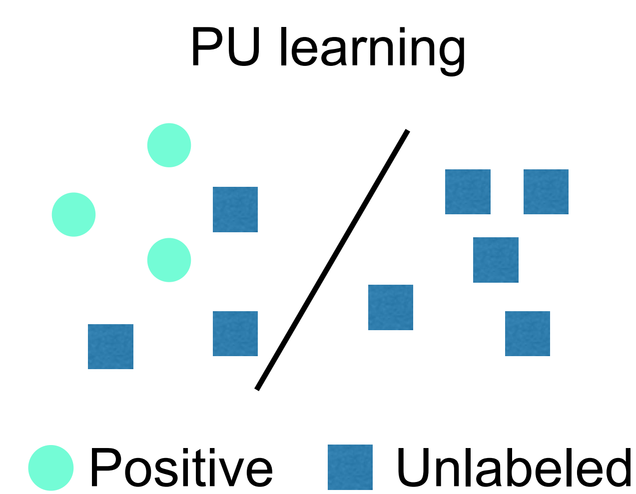
  Entropies of different frequency bands for feature extraction and CNN for classification

- End to end model

  Discovery of iEEG focal without handcraft feature extraction

- Less labels

  Only need a few labelled data by PU learning
  [Prof. Sugiyama's PU algorithms]



EPILEPSYAI
JST CREST

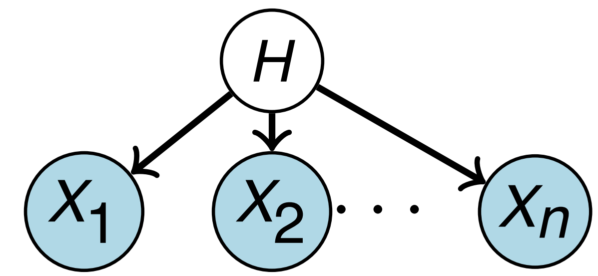
順天堂

Future Work: Tensor Network for Graphical Model

- CP decomposition

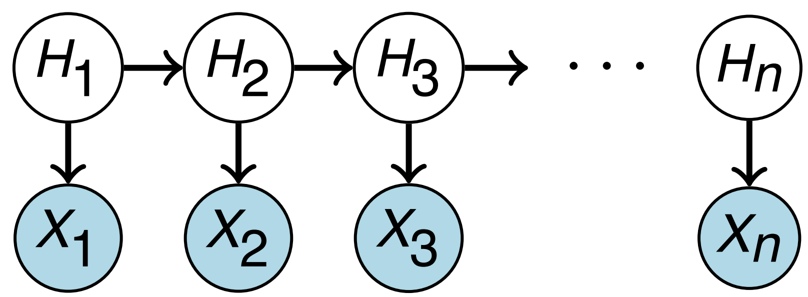
10 variables, 10 states each $\longrightarrow$ 10^{10} entries

$$P(x_1, x_2, x_3, x_4) = \sum_h P(x_1|h)P(x_2|h)P(x_3|h)P(x_4|h)P(h)$$



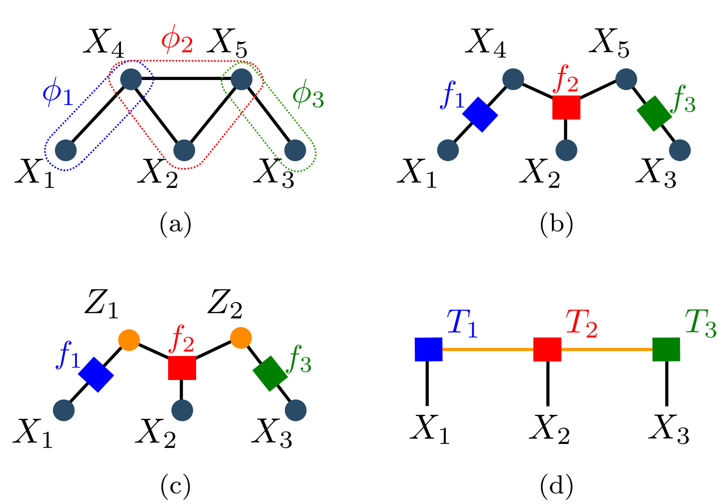
- Markov random field models as tensor train

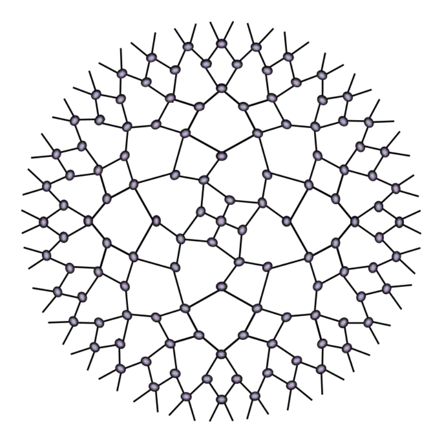
[Novikov et al., ICML 2014]



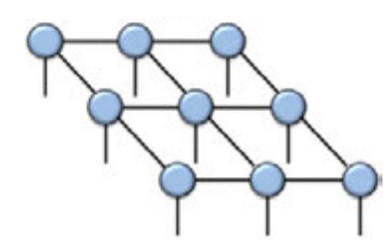
- Undirected graphical model represented as a TT model

[Glasser et al, 2018]





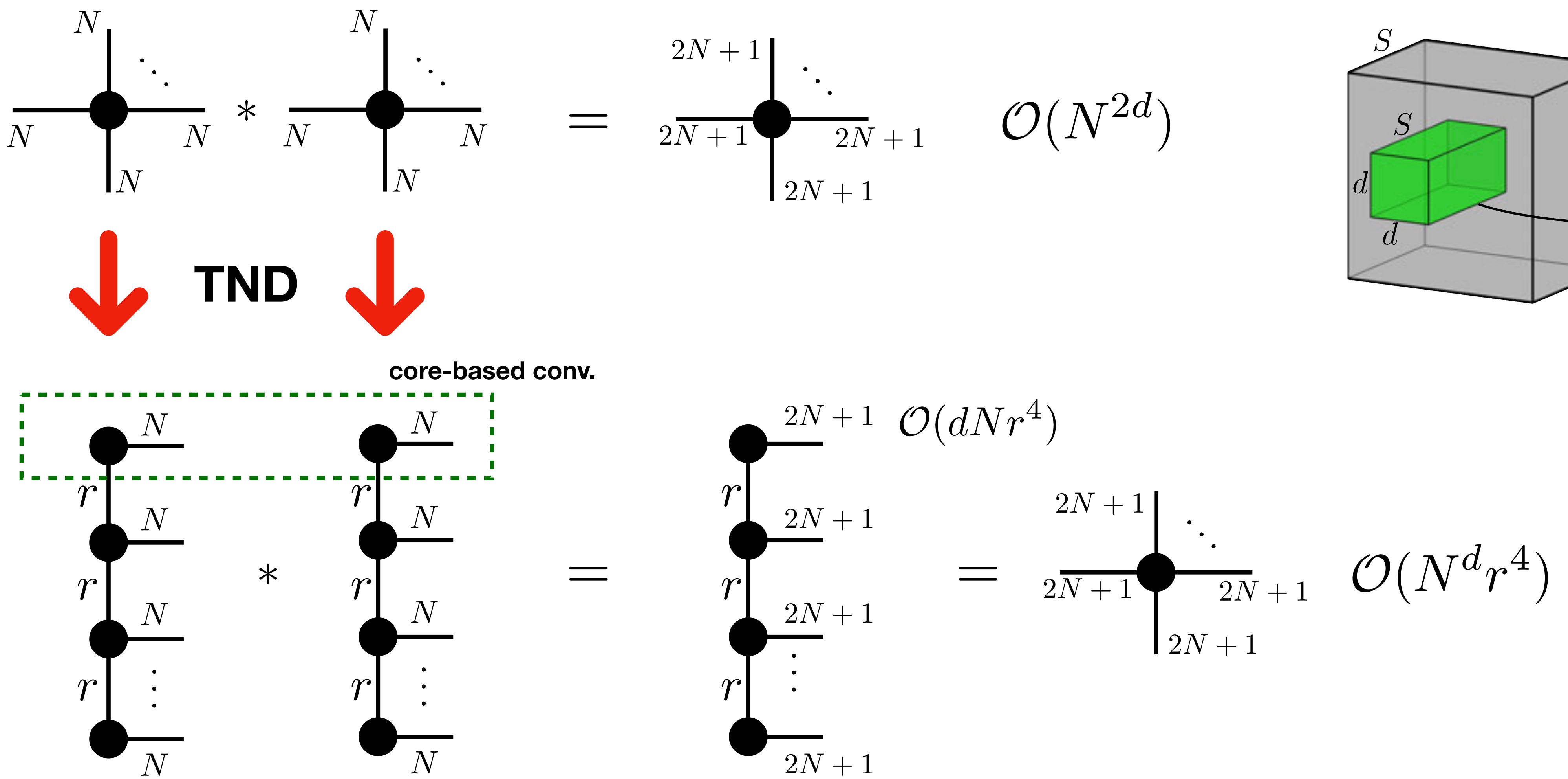
MERA

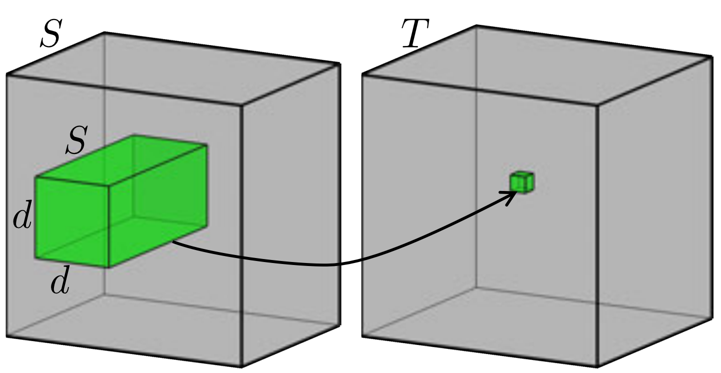


PEPS

Future Work - Acceleration of Tensor Convolution

- High-order convolution (computation and storage)
- Fast convolution via tensor network representation





Collaborations within AIP

- A novel schema for hyper-spectral image restoration

  [He et al., CVPR2019]

Naoto Yokoya

- Dementia detection via tensorizing neural networks

  [Ruikowski et al., NeurIPS 2018 workshop]

Mihoko Otake

- Gene data completion via tensor network

  [Iwata et al., ISMB/ECCB 2019]

Yasuo Tabei

Summary

- Tensor networks are intriguing alternative to traditional machine learning models
- Better scaling, efficient algorithms, opportunities for theoretical insights
- Promising as a framework for machine learning with quantum computing

Achievements in FY2018

Publications (32)

- Conference (19) including AAAI, IJCAI, CVPR, ICASSP, NeurIPS Workshop, ICLR workshop and etc
- Journal (13) including IEEE TNNLS, Signal Processing and etc

Awards

- The 3rd IEEE SPS Japan Best Paper Award
- 2018 SPS Signal Processing Magazine Best Paper