

High-order Tensor Completion via Gradient-based Optimization Under Tensor Train Format

Longhao Yuan^{a,b}, Qibin Zhao^{b,c,*}, Lihua Gui^a, Jianting Cao^{a,b,d,*}

^a*Graduate School of Engineering, Saitama Institute of Technology, Japan*

^b*Tensor Learning Unit, RIKEN Center for Advanced Intelligence Project (AIP), Japan*

^c*School of Automation, Guangdong University of Technology, China*

^d*School of Computer Science and Technology, Hangzhou Dianzi University, China*

Abstract

Tensor train (TT) decomposition has drawn people's attention due to its powerful representation ability and performance stability in high-order tensors. In this paper, we propose a novel approach to recover the missing entries of incomplete data represented by higher-order tensors. We attempt to find the low-rank TT decomposition of the incomplete data which captures the latent features of the whole data and then reconstruct the missing entries. By applying gradient descent algorithms, tensor completion problem is efficiently solved by optimization models. We propose two TT-based algorithms: Tensor Train Weighted Optimization (TT-WOPT) and Tensor Train Stochastic Gradient Descent (TT-SGD) to optimize TT decomposition factors. In addition, a method named Visual Data Tensorization (VDT) is proposed to transform visual data into higher-order tensors, resulting in the performance improvement of our algorithms. The experiments in synthetic data and visual data show high efficiency and performance of our algorithms compared to the state-of-the-art completion algorithms, especially in high-order, high missing rate, and large-scale tensor completion situations.

Keywords: tensor completion, visual data recovery, tensor train decomposition, higher-order tensorization, gradient-based optimization

*Corresponding authors

1. Introduction

Tensors are the high-order generalizations of vectors and matrices. Representing data by tensor can retain the high dimensional form of data and keep adjacent structure information of data. Most of the real-world data are more than two orders. For example, RGB images are order-three tensors ($height \times width \times channel$), videos are order-four tensors ($height \times width \times channel \times time$) and electroencephalography (EEG) signals are order-three tensors ($magnitude \times trails \times time$). When facing data with more than two orders, traditional methods usually transform data into matrices or vectors by concatenation, which leads to spatial redundancy and less efficient factorization[1]. In recent years, many theories, algorithms and applications of tensor methodologies have been studied and proposed [2, 3, 4]. Due to the high compression ability and data representation ability of tensor decomposition, many applications related to tensor decomposition have been proposed in a variety of fields such as image and video completion [5, 6], signal processing [7, 8], brain-computer interface [9], image classification [10], etc.

In practical situations, data missing is ubiquitous due to the error and the noise in data collecting process, resulting in the generation of data outliers and unwanted data entries. Generally, the lynchpin of tensor completion is to find the correlations between the missing entries and the observed entries. Tensor decomposition is to decompose tensor data into decomposition factors which can catch the latent features of the whole data. The basic concept of solving data completion problems by tensor decomposition is that we find the decomposition factors by the partially observed data, then we take advantages of the powerful feature representation ability of the factors to approximate the missing entries. The most studied and classical tensor decomposition models are the CANDECOMP/PARAFAC (CP) decomposition [11, 12], and the Tucker decomposition [13, 14, 15]. CP decomposition decomposes a tensor into a sum of rank-one tensors, and Tucker decomposition approximates a tensor by a core tensor and several factor matrices. There are many proposed tensor completion methods

which employ the two tensor decomposition models. In [5], CP weighted optimization (CP-WOPT) is proposed. It formulates tensor completion problem as a weighted least squares (WLS) problem and uses optimization algorithms to find the optimal CP factors. Fully Bayesian CP Factorization (FBCP) in [6] employs a Bayesian probabilistic model to find the optimal CP factors and CP-rank at the same time. Three algorithms based on nuclear norm minimization are proposed in [16], i.e., SiLRTC, FaLRTC, and HaLRTC. They extend the nuclear norm regularization for matrix completion to tensor completion by minimizing the Tucker rank of the incomplete tensor. In [17], Tucker low- n -rank tensor completion (TLnR) is proposed, and the experiments show better results than the traditional nuclear norm minimization methods.

Though CP and Tucker can obtain relatively high performance in low-order tensors, due to the natural limitations of these two models, when it comes to high-order tensors, the performance of the two decomposition models will decrease rapidly. In recent years, a matrix product state (MPS) model named tensor train (TT) is proposed and becomes popular [18, 19, 20]. For an N th order tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times \dots \times I_N}$, CP decomposition represents data by $\mathcal{O}(\sum_{n=1}^N I_n R)$ model parameters, Tucker model needs $\mathcal{O}(\sum_{n=1}^N I_n R + r^N)$ model parameters, and TT model requires $\mathcal{O}(\sum_{n=1}^N I_n R^2)$ parameters, where R represents the rank of each decomposition model. TT decomposition scales linearly to the tensor order which is the same as CP decomposition. Though the CP model is more compact by ranks, it is difficult to find the optimal CP factors especially when the tensor order is high. Tucker model is more flexible and stable, but model parameters will grow exponentially when the tensor order increases. Tensor train is free from the ‘curse of dimensionality’ so it is a better model to process high-order tensors. In addition to CP-based and Tucker-based tensor completion algorithms, there are several works about TT-based tensor completion. [19] develops the low-TT-rank algorithms for tensor completion. By tensor low-rank assumption based on TT-rank, the nuclear norm regularizations are imposed on the more balanced unfoldings of the tensor, by which the performance improvement is obtained. TT-ALS is proposed in [21], in which the authors employ the

alternative least squares (ALS) method to find the TT decomposition factors to solve tensor completion problem. A gradient-based completion algorithm is discussed in [22], which is to find the TT decomposition by gradient descent method and it shows high performance in high-order tensors and high missing rates tensor completion problems. There are also tensor completion algorithms which are based on the other tensor decomposition models, i.e., tensor ring (TR) decomposition [23, 24] and hierarchical Tucker (HT) decomposition. Based on TR decomposition, works in [25, 26, 27] propose algorithms named TR-ALS, TR-WOPT and TRLRF which apply ALS, gradient descent and nuclear norm minimization methods to solve various tensor completion problems. Moreover, by total variations (TV) and HT decomposition, [28] proposes a completion algorithm named STTC, which explores the global low-rank tensor structure and the local correlation structure of the data simultaneously.

In this paper, we mainly focus on developing efficient tensor completion algorithms based on TT decomposition. Though several tensor completion methods based on TT model have been proposed recently [19, 21, 22], their applicability and effectiveness are limited. The main works of this paper are concluded as follows: 1) Based on optimization methodology and tensor train decomposition, we propose two algorithms named Tensor train Weighted Optimization (TT-WOPT) and Tensor train Stochastic Gradient Descent (TT-SGD) which apply gradient-based optimization algorithms to solve tensor completion problems. 2) We conduct simulation experiments in different tensor orders and compare our algorithms to the state-of-the-art tensor completion algorithms. The superior performance of our algorithms is obtained in both low-order and high-order tensors. 3) We propose a tensorization method named Visual Data Tensorization (VDT) to transform visual data into higher-order tensors, by which the performance of our algorithms is improved. 4) We test the performance of our algorithms on benchmark RGB images, video data, and hyperspectral image data. The higher performance of our algorithms is shown compared to the state-of-the-art algorithms.

The rest of the paper is organized as follows. In Section 2, we state the

notations applied in this paper and introduce the tensor train decomposition. In Section 3, we present the two tensor completion algorithms and analyze the computational complexities of the algorithms. In Section 4, various experiments are conducted on synthetic data and real-world data, in which the proposed algorithms are compared to the state-of-the-art algorithms. We conclude our work in Section 5.

2. Preliminaries and Related works

2.1. Notations

Notations in [2] are adopted in our paper. A scalar is denoted by a normal lowercase/uppercase letter, e.g., $x, X \in \mathbb{R}$, a vector is denoted by a boldface lowercase letter, e.g., $\mathbf{x} \in \mathbb{R}^I$, a matrix is denoted by a boldface capital letter, e.g., $\mathbf{X} \in \mathbb{R}^{I \times J}$, a tensor of order $N \geq 3$ is denoted by an Euler script letter, e.g., $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$.

$\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)}$ denotes a vector sequence, in which $\mathbf{x}^{(n)}$ denotes the n th vector in the sequence. The representations of matrix sequences and tensor sequences are denoted in the same way. An element of tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ of index $\{i_1, i_2, \dots, i_N\}$ is denoted by $x_{i_1 i_2 \dots i_N}$ or $\mathcal{X}(i_1, i_2, \dots, i_N)$. The mode- n matricization (unfolding) of tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ is denoted by $\mathbf{X}_{(n)} \in \mathbb{R}^{I_n \times I_1 \dots I_{n-1} I_{n+1} \dots I_N}$.

Furthermore, the inner product of two tensor \mathcal{X}, \mathcal{Y} with the same size $\mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ is defined as $\langle \mathcal{X}, \mathcal{Y} \rangle = \sum_{i_1} \sum_{i_2} \cdots \sum_{i_N} x_{i_1 i_2 \dots i_N} y_{i_1 i_2 \dots i_N}$. The Frobenius norm of \mathcal{X} is defined by $\|\mathcal{X}\|_F = \sqrt{\langle \mathcal{X}, \mathcal{X} \rangle}$. The Hadamard product is denoted by ‘ $*$ ’ and it is an element-wise product of vectors, matrices or tensors of the same size. For instance, given tensors $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$, $\mathcal{Z} = \mathcal{X} * \mathcal{Y}$, then $\mathcal{Z} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ and $z_{i_1 i_2 \dots i_N} = x_{i_1 i_2 \dots i_N} y_{i_1 i_2 \dots i_N}$ are satisfied. The Kronecker product of two matrices $\mathbf{X} \in \mathbb{R}^{I \times K}$ and $\mathbf{Y} \in \mathbb{R}^{J \times L}$ is $\mathbf{X} \otimes \mathbf{Y} \in \mathbb{R}^{IJ \times KL}$, see more details in [2].

2.2. Tensor Train Decomposition

The most significant feature of TT decomposition is that the number of model parameters will not grow exponentially by the increase of the tensor order. TT decomposition is to decompose a tensor into a sequence of order-three core tensors (factor tensors): $\mathcal{G}^{(1)}, \mathcal{G}^{(2)}, \dots, \mathcal{G}^{(N)}$. The relation between the approximated tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$ and core tensors can be expressed as follow:

$$\mathcal{X} = \ll \mathcal{G}^{(1)}, \mathcal{G}^{(2)}, \dots, \mathcal{G}^{(N)} \gg, \quad (1)$$

where for $n = 1, \dots, N$, $\mathcal{G}^{(n)} \in \mathbb{R}^{R_{n-1} \times I_n \times R_n}$, $R_0 = R_N = 1$, and the notation $\ll \cdot \gg$ is the operation to transform the core tensors to the approximated tensor. It should be noted that, for overall expression convenience, $\mathcal{G}^{(1)} \in \mathbb{R}^{I_1 \times R_1}$ and $\mathcal{G}^{(N)} \in \mathbb{R}^{R_{N-1} \times I_N}$ are considered as two order-two tensors. The sequence R_0, R_1, \dots, R_N is named TT-rank which limits the size of every core tensor. Furthermore, the (i_1, i_2, \dots, i_N) th element of tensor \mathcal{X} can be represented by the multiple product of the corresponding mode-2 slices of the core tensors as:

$$x_{i_1 i_2 \dots i_N} = \prod_{n=1}^N \mathbf{G}_{i_n}^{(n)}, \quad (2)$$

where $\mathbf{G}_{i_1}^{(1)}, \dots, \mathbf{G}_{i_N}^{(N)}$ is the sequence of slices from each core tensor. For $n = 1, 2, \dots, N$, $\mathbf{G}_{i_n}^{(n)} \in \mathbb{R}^{R_{n-1} \times R_n}$ is the mode-2 slice extracted from $\mathcal{G}^{(n)}$ according to each mode of the element index of $x_{i_1 i_2 \dots i_N}$. $\mathbf{G}_{i_1}^{(1)} \in \mathbb{R}^{R_1}$ and $\mathbf{G}_{i_N}^{(N)} \in \mathbb{R}^{R_{N-1}}$ are extracted from first core tensor and last core tensor, they are considered as two order-one matrices for overall expression convenience.

3. Gradient-based Tensor Train Completion

3.1. Tensor train Weighted Optimization (TT-WOPT)

We define $\mathcal{Y} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$ as the partially observed tensor with missing entries and $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$ is the tensor approximated by the core tensors of a TT decomposition. The missing entries of \mathcal{Y} are filled with zero to make \mathcal{Y} to be a real-valued tensor. For modeling the completion problem, the indices

of the missing entries need to be specified. We define a binary tensor $\mathbf{W} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$ named weight tensor in which the indices of missing entries and observed entries of the incomplete tensor \mathbf{Y} can be recorded. Every entry of \mathbf{W} meets:

$$w_{i_1 i_2 \dots i_N} = \begin{cases} 0 & \text{if } y_{i_1 i_2 \dots i_N} \text{ is a missing entry,} \\ 1 & \text{if } y_{i_1 i_2 \dots i_N} \text{ is an observed entry.} \end{cases} \quad (3)$$

The problem of finding the decomposition factors of an incomplete tensor can be formulated by a weight least squares (WLS) model. Define $\mathbf{y}_w = \mathbf{W} * \mathbf{y}$, and $\mathbf{x}_w = \mathbf{W} * \mathbf{x}$, then the WLS model for calculating tensor decomposition factors is formulated by:

$$f(\mathbf{g}^{(1)}, \mathbf{g}^{(2)}, \dots, \mathbf{g}^{(N)}) = \frac{1}{2} \|\mathbf{y}_w - \mathbf{x}_w\|_F^2. \quad (4)$$

This is an optimization objective function w.r.t. all the TT core tensors and we aim to solve the model by gradient descent methods. The relation between the approximated tensor \mathbf{x} and the TT core tensors can be deduced as the following equation [29]:

$$\mathbf{X}_{(n)} = \mathbf{G}_{(2)}^{(n)} (\mathbf{G}_{(1)}^{>n} \otimes \mathbf{G}_{(n)}^{<n}), \quad (5)$$

where for $n = 1, \dots, N$,

$$\mathbf{G}^{>n} = \ll \mathbf{g}^{(n+1)}, \mathbf{g}^{(n+2)}, \dots, \mathbf{g}^{(N)} \gg \in \mathbb{R}^{R_n \times I_{n+1} \times \dots \times I_N}, \quad (6)$$

$$\mathbf{G}^{<n} = \ll \mathbf{g}^{(1)}, \mathbf{g}^{(2)}, \dots, \mathbf{g}^{(n-1)} \gg \in \mathbb{R}^{I_1 \times \dots \times I_{n-1} \times R_{n-1}}. \quad (7)$$

$\mathbf{G}^{>n}$ and $\mathbf{G}^{<n}$ are the tensors generated by merging the selected TT core tensors, and we define $\mathbf{G}^{>N} = \mathbf{G}^{<1} = 1$.

By equation (5), for $n = 1, \dots, N$, the partial derivatives of the objective function (4) w.r.t. the mode-2 matricization of the n th core tensor $\mathbf{g}^{(n)}$ can be inferred as:

$$\frac{\partial f}{\partial \mathbf{G}_{(2)}^{(n)}} = (\mathbf{X}_{w(n)} - \mathbf{Y}_{w(n)}) (\mathbf{G}_{(1)}^{>n} \otimes \mathbf{G}_{(n)}^{<n})^T. \quad (8)$$

After the objective function and the gradients are obtained, we can apply various optimization algorithms to optimize the core tensors. The implementation procedure of TT-WOPT to find the TT decomposition from incomplete tensor \mathbf{Y} is listed in Algorithm 1.

Algorithm 1 Tensor train Weighted Optimization (TT-WOPT)

- 1: **Input:** incomplete tensor \mathcal{Y} , weight tensor \mathcal{W} and TT-rank \mathbf{r} .
 - 2: Randomly initialize the core tensors $\mathcal{G}^{(1)}, \mathcal{G}^{(2)}, \dots, \mathcal{G}^{(N)}$.
 - 3: **While** the optimization stopping condition is not satisfied
 - 4: Compute $\mathcal{X}_w = \mathcal{W} * \ll \mathcal{G}^{(1)}, \mathcal{G}^{(2)}, \dots, \mathcal{G}^{(N)} \gg$.
 - 5: **For** $n=1:N$
 - 6: Compute gradients according to equation (8).
 - 7: **End**
 - 8: Update $\mathcal{G}^{(1)}, \mathcal{G}^{(2)}, \dots, \mathcal{G}^{(N)}$ by gradient descend method.
 - 9: **End while**
 - 10: **Output:** $\mathcal{G}^{(1)}, \mathcal{G}^{(2)}, \dots, \mathcal{G}^{(N)}$.
-

3.2. Tensor Train Stochastic Gradient Descent (TT-SGD)

As seen from equation (4), TT-WOPT computes the gradients by the whole scale of the tensor for every iteration. The computation can be redundant because the missing entries still occupy the computational space. If the scale of data is huge and the number of missing entries is high, then we only need to apply a small amount of the observed entries. In this situation, TT-WOPT can waste much computational storage and the computation will become time-consuming. In order to solve the problems of TT-WOPT as mentioned above, we propose the TT-SGD algorithm which only randomly samples one observed entry to compute the gradients for every iteration.

Stochastic Gradient Descent (SGD) has been applied in matrix and tensor decompositions [30, 31, 32]. For every optimization iteration, we only use one entry which is randomly sampled from the observed entries, and one entry can only influence the gradient of part of the core tensors. For one observed entry of index $\{i_1, i_2, \dots, i_N\}$, if a value approximated by TT core tensors is $x_{i_1 i_2 \dots i_N}$ and the observed value (real value) is $y_{i_1 i_2 \dots i_N}$, by considering equation (2), the objective function can be formulated by:

$$f(\mathbf{G}_{i_1}^{(1)}, \mathbf{G}_{i_2}^{(2)}, \dots, \mathbf{G}_{i_N}^{(N)}) = \frac{1}{2} \left\| y_{i_1 i_2 \dots i_N} - \prod_{k=1}^N \mathbf{G}_{i_k}^{(k)} \right\|_F^2. \quad (9)$$

For $n = 1, 2, \dots, N$, the partial derivatives of every corresponding slice $\mathbf{G}_{i_n}^{(n)}$

w.r.t. index $\{i_1, i_2, \dots, i_N\}$ is calculated as:

$$\frac{\partial f}{\partial \mathbf{G}_{i_n}^{(n)}} = (x_{i_1 i_2 \dots i_N} - y_{i_1 i_2 \dots i_N}) \left(\prod_{k=n+1}^N \mathbf{G}_{i_k}^{(k)} \prod_{k=1}^{n-1} \mathbf{G}_{i_k}^{(k)} \right)^T. \quad (10)$$

From the equation we can see, the computational complexity of TT-SGD is not related to the scale of the observed tensor or the number of observed entries, so it can process large-scale data by much smaller computational complexity than TT-WOPT. This algorithm is also suitable for online/real-time learning. The optimization process of TT-SGD is listed in Algorithm 2:

Algorithm 2 Tensor Train Stochastic Gradient Descent (TT-SGD)

- 1: **Input:** incomplete tensor \mathcal{Y} and $TT - rank \mathbf{r}$.
 - 2: Randomly initialize core tensors $\mathbf{G}^{(1)}, \mathbf{G}^{(2)}, \dots, \mathbf{G}^{(N)}$.
 - 3: **While** the optimization stopping condition is not satisfied
 - 4: Randomly sample $y_{i_1 i_2 \dots i_N}$ from \mathcal{Y} .
 - 5: **For** $n=1:N$
 - 6: Compute the gradients of the core tensors by equation (10).
 - 7: **End**
 - 8: Update $\mathbf{G}_{i_1}^{(1)}, \mathbf{G}_{i_2}^{(2)}, \dots, \mathbf{G}_{i_N}^{(N)}$ by gradient descent method.
 - 9: **End while**
 - 10: **Output:** $\mathbf{G}^{(1)}, \mathbf{G}^{(2)}, \dots, \mathbf{G}^{(N)}$.
-

3.3. Computational Complexity

For tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$, we assume all I_1, I_2, \dots, I_N is equal to I , and $R_1 = R_2 = \dots = R_{N-1} = R$. According to equation (8) and (10), the time complexity of TT-WOPT and TT-SGD are $\mathcal{O}(NI^N + NI^{N-1}R^2)$ and $\mathcal{O}(N^2R^3)$ respectively, and the space complexity of the two algorithms is $\mathcal{O}(I^N + I^{N-1}R^2)$ and $\mathcal{O}(R^2)$ respectively. Though TT-WOPT has larger computational complexity, it has a steady and fast convergence when processing normal-size data. TT-SGD is free from data dimensionality and the complexity of every iteration is extremely low, so it is more suitable to process large-scale data. It should be noted that for every iteration of TT-SGD, we can also apply the batch-based

SGD method which calculates the summation of the gradients of batch-sized entries for every iteration. Though this can improve the stability of TT-SGD and the algorithm might need fewer iterations to be converged, the computational complexity will be increased and more computational time is needed for every iteration. In this paper, we only apply batch-one SGD algorithm, and the synthetic experiment in the next section show that our method can also achieve fast and stable convergence. The code of the proposed algorithms is available at https://github.com/yuanlonghao/T3C_tensor_completion.

4. Experiment results

In this section, simulation experiments are conducted to show the performance of our algorithms and the compared algorithms under various tensor orders. For real-world data experiments, we test our algorithms by color images, video data and hyperspectral image data. TT-WOPT and TT-SGD are compared with several state-of-the-art algorithms: TT-ALS [21], SiLRTC-TT [19], TRALS [25], STTC [28], CP-WOPT [5], FBCP [6], HaLRTC and FaLRTC [16], and TLnR [17]. For all the compared algorithms, the input incomplete tensor is $\mathcal{W} * \mathcal{Y}$, where \mathcal{Y} is the fully observed true tensor, \mathcal{W} is the binary tensor recording the position of observed entries. The final completed tensor \mathcal{Z} is calculated by $\mathcal{Z} = (1 - \mathcal{W}) * \mathcal{X} + \mathcal{W} * \mathcal{Y}$, where \mathcal{X} is the output tensor obtained by each algorithm. We apply relative squared error (RSE) which is defined as $RSE = \|\mathcal{Y} - \mathcal{Z}\|_F / \|\mathcal{Y}\|_F$ to evaluate the completion performance for each algorithm. For experiments of random missing cases, we randomly remove data points according to different missing rates m_r , which is defined as $m_r = 1 - M / \prod_{n=1}^N I_n$, where M is the number of the observed entries. Moreover, to evaluate the completion quality of visual data, we introduce PSNR (Peak Signal-to-noise Ratio). PSNR is obtained by $PSNR = 10 \log_{10}(255^2 / MSE)$, where MSE is deduced by $MSE = \|\mathcal{Z} - \mathcal{Y}\|_F^2 / num(\mathcal{Z})$, and $num(\cdot)$ denotes the number of the element of the tensor.

For optimization method of TT-WOPT, in order to have a clear comparison

with CP-WOPT which is also based on gradient descent methods, we adopt the same optimization method as paper [5]. The paper applies nonlinear conjugate gradient (NCG) with Hestenes-Stiefel updates [33] and the Moré-Thuente line search method [34]. The optimization method is implemented by an optimization toolbox named Pablano Toolbox [35]. For TT-SGD, we employ an algorithm named Adaptive Moment Estimation (Adam) as our gradient descent method, it has prominent performance on stochastic-gradient-based optimization [36, 37]. The update rule of Adam is as follow:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{v_t + \epsilon}} m_t, \quad (11)$$

where t is the iteration time of optimization value θ , η and ϵ are hyper parameters, m_t and v_t are the first moment estimate and second moment estimate of gradient g_t respectively. $m_t = \beta_1 m_{t-1} + (1 - \beta_1)g_t$, $v_t = \beta_2 v_{t-1} + (1 - \beta_2)g_t^2$, where β_1 and β_2 are hyper parameters. For choosing the hyper parameters in Adam method, we adopt the reference values from paper [36]. The values of β_1 , β_2 and ϵ are set as 0.9, 0.999 and 10^{-8} respectively. The selection of learning rate is essential to the convergence speed and the performance of the gradient-based algorithms, in our experiments, we empirically choose the learning rate η from $\{0.0001, 0.0005, 0.001\}$ to obtain the best convergence speed and the best performance. In addition, all the data in our experiments are regularized to 0 to 1 to make the algorithms more effective.

We mainly adopt two optimization stopping conditions for all the compared completion algorithms. One is the error of two adjacent iterations of the objective function value: $|f_t - f_{t-1}| \leq tol$, where f_t is the objective function value of the t th iteration and we set $tol = 1e-4$ in our experiment. The other stopping condition is the maximum number of iteration which is set according to the scale of data and different algorithms, e.g., the maximum iteration for most algorithms are set as 500 and for TT-SGD it usually set from 10^5 to 10^7 . If one of the two conditions is satisfied, the optimization will be stopped. All the computations are conducted on a Mac PC with Intel Core i7 and 16GB DDR3 memory, and the computational time of the algorithms are recorded in some

experiments based on this configuration.

4.1. Synthetic Data

We apply synthetic data generated from a highly oscillating function: $f(x) = \sin\frac{x}{4}\cos(x^2)$ [38] in our simulation experiments. The synthetic data is expected to be well approximated by tensor decomposition models. We sample I^N entries from the values generated from the function, then the sampled values are reshaped to the desired tensor size. We employ four different tensor structures: $26 \times 26 \times 26$ (3D), $7 \times 7 \times 7 \times 7 \times 7$ (5D), $4 \times 4 \times 4 \times 4 \times 4 \times 4 \times 4$ (7D), and $3 \times 3 \times 3 \times 3 \times 3 \times 3 \times 3 \times 3$ (9D), then we test TT-SGD, TT-WOPT, TT-ALS, SiLRTC-TT, TR-ALS, CP-WOPT and HaLRTC on the synthetic data. For parameter settings, the hyper-parameters of each algorithm are tuned to obtain the best performance. For simplicity, we set values of each TT-rank and TR-rank identically, i.e., $R_1 = \dots = R_{N-1}$ for TT and $R_1 = \dots = R_N$ for TR. Moreover, the TT-rank, TR-rank and CP-rank are set as 12, 10 and 30 under all the different tensor orders for the corresponding algorithms to make a clear comparison of the completion performance. In addition, the maximum iteration of TT-SGD is set as 10^5 , and iteration for other algorithms are all set as 500.

The graphs of Figure 1 show the experiment results of RSE values, which change by different m_r (from 0.1 to 0.9) under the four different tensor orders. From the figure, we can see that TT-WOPT and TT-SGD show high performance in all the cases. HaLRTC only shows high performance in 3D tensor case, and CP-WOPT and SiLRTC show stable but low performance in every case. Though TT-ALS and TR-ALS show higher performance than our algorithms in some low missing rate cases, the drastic performance decrease can be obtained from them when the missing rate increases, and our algorithms always show high and stable performance.

For the next synthetic data experiment, we aim to look into the convergence performance of the proposed TT-SGD. The four tensors which applied in the previous experiment is employed as the input data. We record the value of loss function (i.e., $\frac{1}{2}\|\mathcal{Z} - \mathcal{Y}\|_F^2$) for every 10^3 iterations and Figure 2 shows the con-

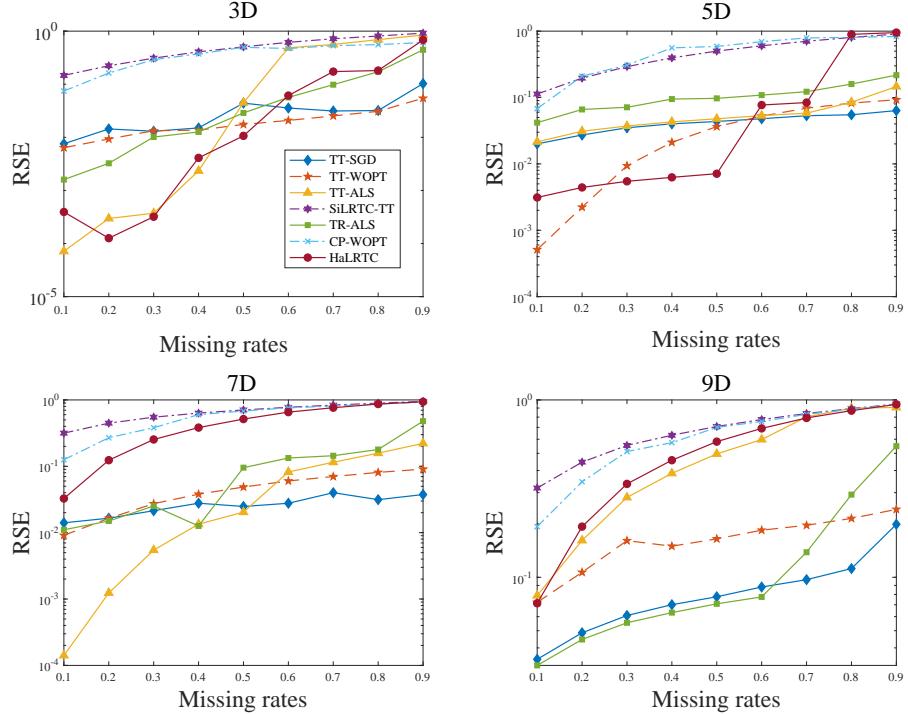


Figure 1: RSE comparison of seven algorithms under four different tensor orders. The missing rate is tested from 0.1 to 0.9.

vergence status of TT-SGD when the missing rate is 0.1, 0.5 and 0.9 respectively. Though our TT-SGD needs large numbers of iteration to be converged, the computational complexity of each iteration is rather low (i.e., N^2R^3), and only one entry is sampled to calculate the gradient for every iteration. For TT-SGD, the running time of reaching 10^5 iterations for the 3D, 5D, 7D, 9D data under the parameter setting in the experiment is 10.09 seconds, 25.09 seconds, 45.86 seconds and 75.41 seconds respectively, while for TT-WOPT, it takes about two times longer than TT-SGD (i.e., 18.80 seconds, 41.84 seconds, 100.02 seconds and 122.77 seconds) to converge to the same RSE values. The performance and computation time manifest the effectiveness of the TT-SGD algorithm.

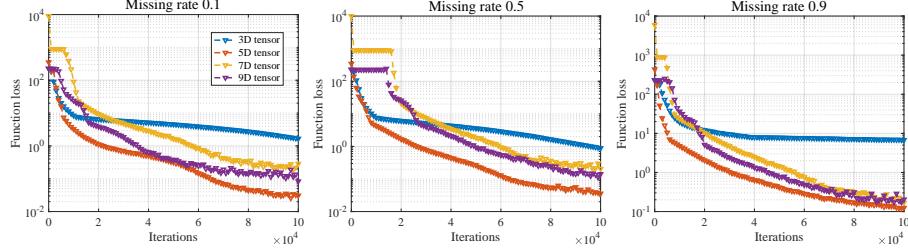


Figure 2: Convergence performance of TT-SGD under four different synthetic tensors. From left to right, the missing rate of the data in each figure is 0.1, 0.5 and 0.9 respectively.

4.2. Visual Data Tensorization (VDT) method

From the simulation results we can see, our proposed algorithms achieve high and stable performance in high-order tensors. In this section, we provide a Visual Data Tensorization (VDT) method to transform low-order tensor into higher-order tensor and improve the performance of our algorithms. The VDT method is derived from an image compression and entanglement methodology [39] which is to transform a gray-scale image of size $2^l \times 2^l$ into a real ket of a Hilbert space. The method cast the image to a higher-order tensor structure with an appropriate block structured addressing. Similar method named KA augmentation is proposed in [19] which extends the method in [39] to order-three visual data of size $2^l \times 2^l \times 3$. Our VDT method is a generalization of the KA augmentation, and the visual data of various data sizes can be applied to our tensorization method. For visual data like RGB image, video, hyperspectral image, the first two orders of the tensor (e.g., $\mathbf{Y} \in \mathbb{R}^{U \times V}$) are named as the image modes. The 2D representation of the image modes cannot fully exploit the correlation and local structure of the data, so we propose the VDT method to strengthen the local structure correlation of visual data. The VDT method operates as follows: if the first two orders of a visual data tensor is $U \times V$ and can be reshaped to $u_1 \times u_2 \times \dots \times u_l \times v_1 \times v_2 \times \dots \times v_l$, then VDT method permutes and reshapes the data to size $u_1 v_1 \times u_2 v_2 \times \dots \times u_l v_l$ and obtain the higher-order

representation of the visual data. This higher-order tensor is a new structure of the original data: the first order of this higher-order tensor corresponds to a $u_1 \times v_1$ pixel block of the image, and the following orders of u_2v_2, \dots, u_lv_l describe the expanding larger-scale partition of the image. Based on VDT method, TT-based algorithms can efficiently exploit the structure information of visual data and achieve a better low-rank representation. After the tensorized data is calculated by the completion algorithms, a reverse operation of VDT is conducted to get the original image structure. The diagrams to explain the procedure of VDT are shown in Figure 3.

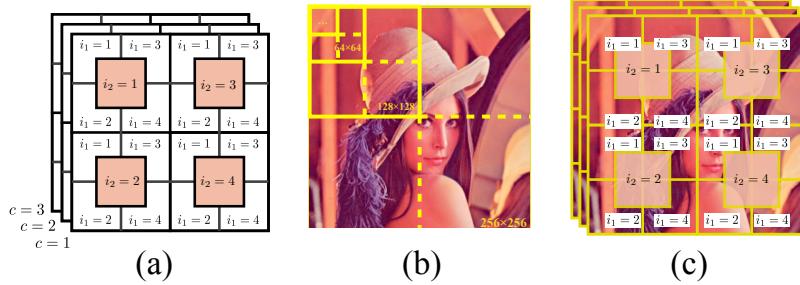


Figure 3: Illustration of the proposed VDT method. Figure (a) is the example of applying VDT method on an $I \times I \times C$ tensor. Figure (b) and Figure (c) shows the example of the VDT operation on a $256 \times 256 \times 3$ image.

To verify the effectiveness of our VDT method, we choose a benchmark image ‘Lena’ with 0.9 missing rate. We compare the performance of the six algorithms (TT-WOPT, TT-SGD, CP-WOPT, FBCP, HaLRTC and TLnR) under three different data structures: order-three tensor, order-nine tensor without VDT, order-nine tensor generated by VDT method. The order-three tensor applies original image data structure of size $256 \times 256 \times 3$. The nine-order tensor without VDT is generated by directly reshaping data to the size $4 \times 4 \times 3$. For nine-order tensor with VDT method, firstly the original data is reshaped to a order-seventeen tensor of size $2 \times 2 \times 3$ and then it is permuted according to the order of $\{1\ 9\ 2\ 10\ 3\ 11\ 4\ 12\ 5\ 13\ 6\ 14\ 7\ 15\ 8\ 16\ 17\}$. Finally we reshape

the tensor to a nine-order tensor of size $4 \times 4 \times 3$. This nine-order tensor with VDT is considered to be a better structure of the image data. The first order of the nine way tensor contains the data of a 2×2 pixel block of the image and the following orders of the tensor describe the expanding pixel blocks of the image. Most of the parameter settings follow the previous synthetic data experiments, and we tune the TT-rank, CP-rank and Tucker-rank of the corresponding algorithms to obtain the best performance. Figure 4 and Table 1 show the visual results and numerical results of the six algorithms under the three different data structure. We can see that in the three-order tensor case, the results among the algorithms are similar. However, for nine-order cases, other algorithms fail the completion task while TT-WOPT and TT-SGD perform well. Furthermore, when the image is transformed to nine-order tensor by VDT method, we see the distinct improvement of our two algorithms.

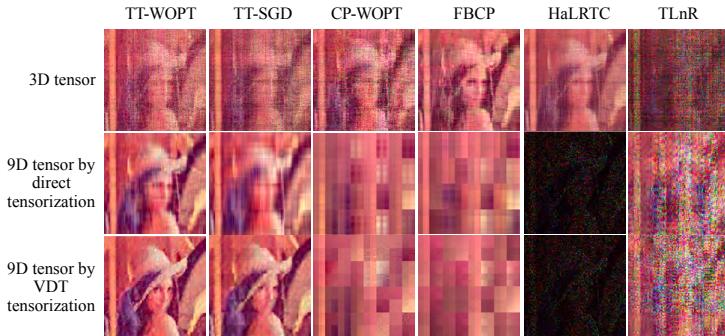


Figure 4: Visual results for completion of the 0.9 random missing ‘Lena’ image under six algorithms. The first row applies original order-three tensor data, the second row applies order-nine tensor data without VDT method, and the third row applies order-nine tensor data generated by VDT method.

4.3. Benchmark Image Completion

From the previous experiments we can see, TT-based and TR-based algorithms can be applied to higher-order tensors, and significant improvement of

Table 1: Numerical results of completion performance (RSE and PSNR) of six algorithms under three tensor structures of image ‘Lena’.

		TT-WOPT	TT-SGD	CP-WOPT	FBCP	HaLRTC	TLnR
three-order	RSE	0.2822	0.2604	0.3392	0.1942	0.1981	0.6552
	PSNR	16.12	16.84	14.53	19.36	19.18	8.802
nine-order	RSE	0.1558	0.1793	0.2562	0.2682	0.9310	1.207
	PSNR	21.31	20.06	16.95	16.57	5.746	3.486
nine-order VDT	RSE	0.1262	0.1493	0.2573	0.2687	0.9301	0.7114
	PSNR	23.21	21.77	16.97	16.57	5.751	10.84

TT-based algorithms can be seen when the VDT method is applied to the image tensorization. However, for algorithms which are based on CP decomposition and Tucker decomposition, higher-order tensorization will decrease the performance. In later experiments, we only apply the VDT method to TT-WOPT, TT-SGD, TT-ALS, SILRTC-TT and TR-ALS. For CP-WOPT, FBCP, TLnR, STTC and HaLRTC, we keep the original data structure to get better results.

In this experiment, we consider several irregular missing cases (the scratch missing, the whole row missing and the block missing) and some high-random-missing cases on benchmark RGB images. The parameter settings for each compared algorithms are tuned to get the best performance. The completion results from Figure 5 and Table 2 we can see, our algorithms show high completion performance in all the missing cases. Moreover, for irregular missing cases and 0.8 random missing cases, STTC and HaLRTC performs well and achieve low RSE values. However, the two algorithms fail to solve the completion task when the random missing rate is 0.9 and 0.99, this is because the nuclear-norm-based and total-variations-based algorithms cannot explore low-rank and local information when only a very small amount of entries is obtained. It should be noted that the 0.99 random missing case is a challenging task among all the image completion algorithms. Our two proposed algorithms with VDT method can achieve high performance under this situation while the other algorithms fail.

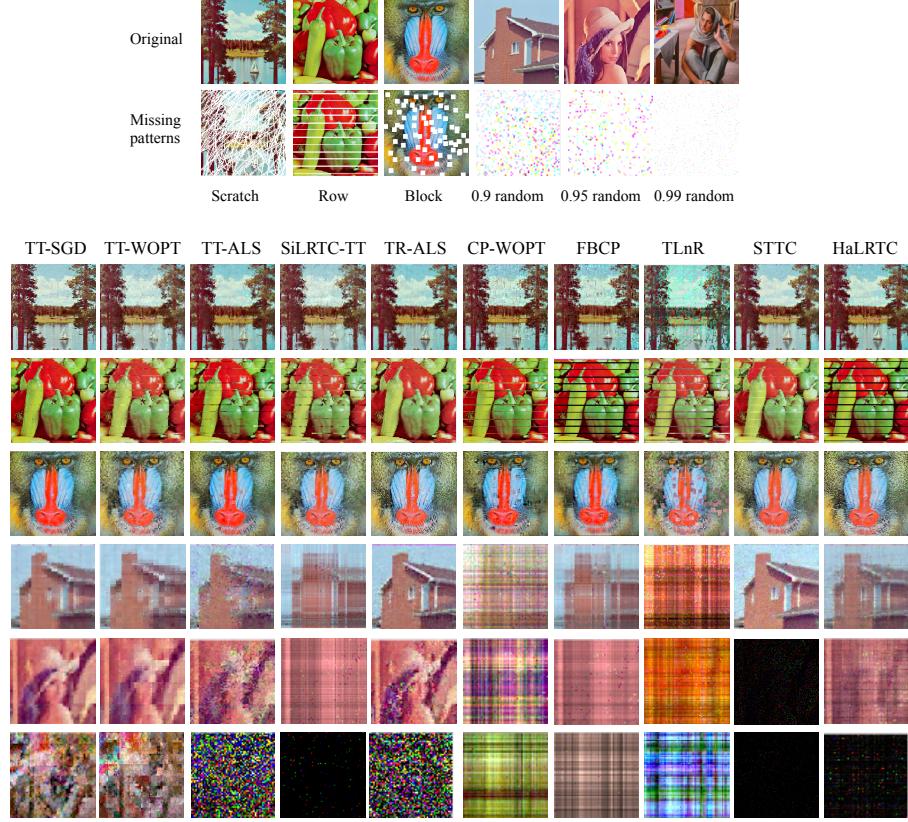


Figure 5: The first and second row of the figure is the fully observed benchmark images and the corresponding missing patterns respectively, below which the visual completion results of the ten algorithms under the different missing patterns (i.e., scratch missing, row missing, block missing, 0.9 random missing, 0.95 random missing, and 0.99 random missing) are shown.

4.4. Video and Hyperspectral Image Completion

For large-scale data completion task, we test a video and a hyperspectral image (HSI) in the following experiments. For our proposed algorithms, we only test TT-SGD because TT-SGD is better for large-scale data than TT-WOPT. In addition, when large-scale data is employed, many algorithms which work well

Table 2: Comparison of the inpainting performance (RSE and PSNR) of ten algorithms under six missing situations.

missing patterns	indices	TT-SGD	TT-WOPT	TT-ALS	SILRTC-TT	TR-ALS	CP-WOPT	FBCP	TLnR	STTC	HaLRTC
Scratch	RSE	0.09946	0.1455	0.1319	0.1522	0.1173	0.2160	0.1185	0.2871	0.1085	0.1168
	PSNR	24.93	21.62	22.48	21.23	23.49	18.19	23.40	15.72	24.17	23.53
Row	RSE	0.07319	0.09629	0.1385	0.1379	0.07325	0.1653	0.3605	0.1797	0.1069	0.3605
	PSNR	27.91	25.54	22.38	22.41	27.91	20.84	14.07	20.12	24.62	14.07
Block	RSE	0.08084	0.09196	0.09671	0.09511	0.08517	0.1391	0.1147	0.1579	0.07315	0.08167
	PSNR	27.20	26.09	25.65	29.86	26.75	22.49	24.17	21.39	28.07	27.12
0.9 random	RSE	0.1444	0.1635	0.1891	0.1969	0.1090	0.3209	0.1967	0.5815	0.1845	0.1621
	PSNR	21.11	20.03	18.77	18.41	23.55	14.17	18.43	9.01	18.98	20.11
0.95 random	RSE	0.1576	0.1797	0.2547	0.2865	0.2147	0.4045	0.2850	0.5557	-	0.2820
	PSNR	21.17	20.32	17.00	15.98	18.49	12.98	16.02	10.23	-	16.11
0.99 random	RSE	0.3318	0.2520	-	0.4049	-	0.4749	0.4074	0.8545	-	0.9129
	PSNR	15.81	15.30	-	6.98	-	12.70	14.03	7.30	-	7.03

on benchmark images will become inefficient or ineffective, so we compare TT-SGD to only several algorithms (TT-ALS, CP-WOPT, FBCP, and HaLRTC).

First, we test a video which records a moving train. The size of the data is $320 \times 256 \times 3 \times 100$ and the background of the video changes by frames. By the VDT method, we first reshape the data to size $2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 5 \times 2 \times 2 \times 2 \times 2 \times 2 \times 4 \times 3 \times 100$, then permute it by index $\{1\ 8\ 2\ 9\ 3\ 10\ 4\ 11\ 5\ 12\ 6\ 13\ 7\ 14\ 15\ 16\}$, and finally we reshape it to size $4 \times 4 \times 4 \times 4 \times 4 \times 4 \times 20 \times 3 \times 100$ as the input tensor. We compare three random missing cases ($m_r = 0.7$, $m_r = 0.9$ and $m_r = 0.99$) in this experiment. Part of the visual results are shown in Figure 6, and the numerical results are shown in Table 3. The performance of TT-SGD outperforms other compared algorithms. More specifically, it can recover the video well even there is only 1% sampled entries while other compared algorithms fail in this high missing rate case. It should also be noted that the time cost of TT-SGD is lower than the other compared algorithms, which shows high efficiency of TT-SGD.

Then we test TT-SGD, CP-WOPT, FBCP and HaLRTC on a hyperspectral image (HSI) of size $256 \times 256 \times 191$ recorded by a satellite. Due to the inferior working condition of satellite sensors, the collected data often has Gaussian noise, impulse noise, dead lines, and stripes [41]. In this experiment, we first consider the situation when the HSI has ‘dead lines’, which is a common missing

Table 3: Numerical results (RSE and PSNR) on video completion experiments of five algorithms under three random missing cases.

	$m_r = 0.7$			$m_r = 0.9$			$m_r = 0.99$		
Algorithm	RSE	PSNR	Time	RSE	PSNR	Time	RSE	PSNR	Time
TT-SGD	0.1459	22.67	680.94	0.2045	19.87	674.17	0.2185	19.24	698.11
TT-ALS	0.2116	19.48	7100.43	0.2400	18.39	1622.42	0.2557	17.8466	793.76
CP-WOPT	0.2673	17.41	825.06	0.3264	15.67	790.60	0.3610	14.80	814.44
FBCP	0.2204	19.11	870.89	0.2547	17.86	920.78	0.3258	15.72	720.01
HaLRTC	0.1758	21.16	1132.05	0.2562	17.78	1044.88	0.8844	7.016	1121.37

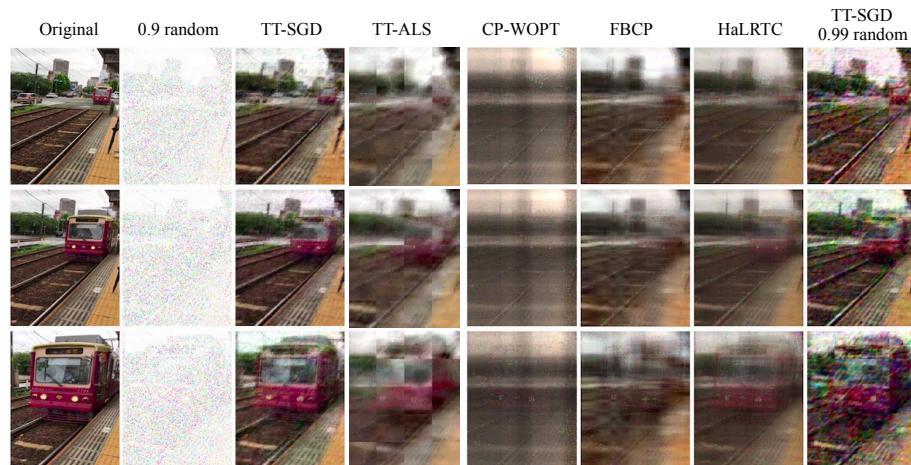


Figure 6: Video completion results of TT-SGD, TT-ALS, CP-WOPT, FBCP, and HaLRTC under random missing cases. The first row to the last row show the completion results of the 1st frame, the 75th frame and the 100th frame of the video respectively.

case in HSI record. Then we consider the case when only 1% of the data is obtained, which is meaningful in data compression and transformation. We transform the HSI data to $16 \times 16 \times 16 \times 16 \times 191$ by VDT method as the input for TT-SGD and apply original three-order tensor as the input for the other compared algorithms. We set TT-ranks as 48 and 24 for dead line missing case and 99% missing case respectively. The visual completion results in Figure 7 shows the image of the first channel of the HSI and the numerical results are the evaluation of the overall completion performance.

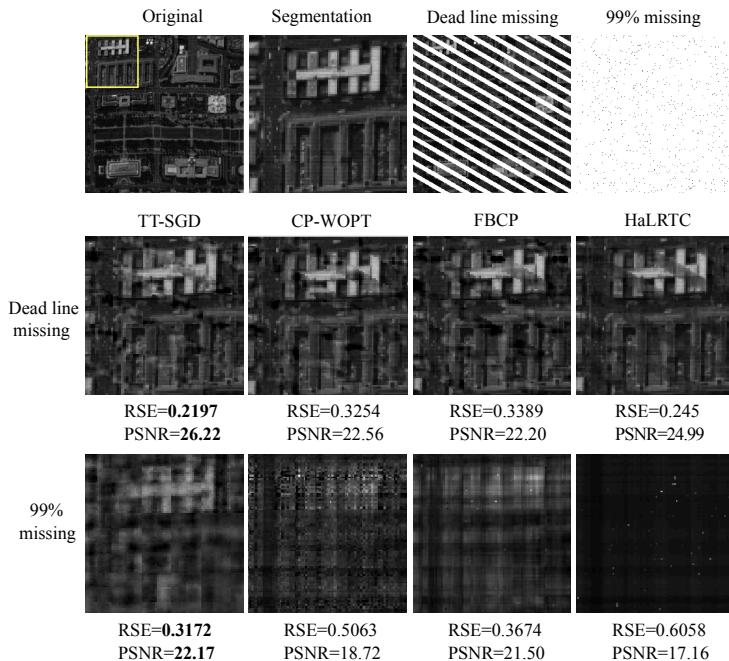


Figure 7: HSI completion results of the four algorithms. We show the image of the first channel of the HSI. The first row is the original image, the segmentation to show the completion performance, the dead line missing pattern, and the 0.99 random missing pattern. The second row and the third row show the completion results.

TT-SGD performs best among the algorithms at both dead line missing case and 99% random missing case. In 99% random missing case, HaLRTC fails the completion task, while CP-WOPT and FBCP obtain lower performance than

TT-SGD. In addition, it should be noted that the volume of data is about 1.25×10^7 , and when the iteration reaches 1×10^6 (16% of the total data), the optimization of TT-SGD is converged. This indicates that TT-SGD has fast and efficient computation.

5. Conclusion

In this paper, in order to solve the tensor completion problem, based on tensor train decomposition and gradient descent method, we propose two tensor completion algorithms named TT-WOPT and TT-SGD. We first cast the completion problem into solving the optimization models, then we use gradient descent methods to find the optimal core tensors of TT decomposition. Finally, the TT core tensors are applied to approximate the missing entries of the incomplete tensor. Furthermore, to improve the performance of the proposed algorithms, we propose the VDT method to tensorize visual data to higher-order. We conduct simulation experiments and visual data experiments to compare our algorithms to the state-of-the-art algorithms. From the simulation experiments we can see, the performance of our algorithms stays stable when the tensor order increases. Moreover, the visual data experiments show that after higher-order tensorization by VDT, the performance of our two algorithms can be improved. Our algorithms outperform the compared state-of-the-art algorithms in various missing situations, particularly when the tensor order is high and the missing rate is high. More specially, our algorithms with VDT method can process extreme high random missing situation (i.e., 99% random missing) well while other algorithms fail. Besides, our proposed TT-SGD achieves low computational complexity and high efficiency in processing large-scale data.

The high performance of the proposed algorithms shows that TT-based tensor completion is a promising aspect. It should be noted that TT-rank setting is essential to obtain better experiment results and it is selected manually in common. We will extend our algorithms by choosing TT-rank automatically in our future work.

Acknowledgement

This work was supported by JSPS KAKENHI (Grant No. 17K00326, 15H04002, 18K04178), JST CREST (Grant No. JPMJCR1784) and the National Natural Science Foundation of China (Grant No. 61773129).

References

References

- [1] A. Shashua, T. Hazan, Non-negative tensor factorization with applications to statistics and computer vision, in: Proceedings of the 22nd international conference on Machine learning, ACM, 2005, pp. 792–799.
- [2] T. G. Kolda, B. W. Bader, Tensor decompositions and applications, SIAM review 51 (3) (2009) 455–500.
- [3] M. A. O. Vasilescu, D. Terzopoulos, Multilinear subspace analysis of image ensembles, in: Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on, Vol. 2, IEEE, 2003, pp. II–93.
- [4] T. Franz, A. Schultz, S. Sizov, S. Staab, Triplerank: Ranking semantic web data by tensor decomposition, The Semantic Web-ISWC 2009 (2009) 213–228.
- [5] E. Acar, D. M. Dunlavy, T. G. Kolda, M. Mørup, Scalable tensor factorizations for incomplete data, Chemometrics and Intelligent Laboratory Systems 106 (1) (2011) 41–56.
- [6] Q. Zhao, L. Zhang, A. Cichocki, Bayesian cp factorization of incomplete tensors with automatic rank determination, IEEE transactions on pattern analysis and machine intelligence 37 (9) (2015) 1751–1763.
- [7] L. De Lathauwer, J. Castaing, Blind identification of underdetermined mixtures by simultaneous matrix diagonalization, IEEE Transactions on Signal Processing 56 (3) (2008) 1096–1105.

- [8] D. Muti, S. Bourennane, Multidimensional filtering based on a tensor approach, *Signal Processing* 85 (12) (2005) 2338–2353.
- [9] J. Mocks, Topographic components model for event-related potentials and some biophysical considerations, *IEEE transactions on biomedical engineering* 35 (6) (1988) 482–484.
- [10] A. Shashua, A. Levin, Linear image coding for regression and classification using the tensor-rank principle, in: *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, Vol. 1, IEEE, 2001, pp. I–I.
- [11] L. Sorber, M. Van Barel, L. De Lathauwer, Optimization-based algorithms for tensor decompositions: Canonical polyadic decomposition, decomposition in rank-($l_r, l_r, 1$) terms, and a new generalization, *SIAM Journal on Optimization* 23 (2) (2013) 695–720.
- [12] J. H. d. M. Goulart, M. Boizard, R. Boyer, G. Favier, P. Comon, Tensor cp decomposition with structured factor matrices: Algorithms and performance, *IEEE Journal of Selected Topics in Signal Processing* 10 (4) (2016) 757–769.
- [13] L. R. Tucker, Some mathematical notes on three-mode factor analysis, *Psychometrika* 31 (3) (1966) 279–311.
- [14] L. De Lathauwer, B. De Moor, J. Vandewalle, On the best rank-1 and rank- (r_1, r_2, \dots, r_n) approximation of higher-order tensors, *SIAM journal on Matrix Analysis and Applications* 21 (4) (2000) 1324–1342.
- [15] C.-Y. Tsai, A. M. Saxe, D. Cox, Tensor switching networks, in: *Advances in Neural Information Processing Systems*, 2016, pp. 2038–2046.
- [16] J. Liu, P. Musalski, P. Wonka, J. Ye, Tensor completion for estimating missing values in visual data, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35 (1) (2013) 208–220.

- [17] M. Filipović, A. Jukić, Tucker factorization with missing data with application to low-n-rank tensor completion, *Multidimensional systems and signal processing* 26 (3) (2015) 677–692.
- [18] I. V. Oseledets, Tensor-train decomposition, *SIAM Journal on Scientific Computing* 33 (5) (2011) 2295–2317.
- [19] J. A. Bengua, H. N. Phien, H. D. Tuan, M. N. Do, Efficient tensor completion for color image and video recovery: Low-rank tensor train, *IEEE Transactions on Image Processing* 26 (5) (2017) 2466–2479.
- [20] Y. Yang, D. Krompass, V. Tresp, Tensor-train recurrent neural networks for video classification, in: *International Conference on Machine Learning*, 2017, pp. 3891–3900.
- [21] W. Wang, V. Aggarwal, S. Aeron, Tensor completion by alternating minimization under the tensor train (tt) model, arXiv preprint arXiv:1609.05587.
- [22] L. Yuan, Q. Zhao, J. Cao, Completion of high order tensor data with missing entries via tensor-train decomposition, in: *International Conference on Neural Information Processing*, Springer, 2017, pp. 222–229.
- [23] Q. Zhao, G. Zhou, S. Xie, L. Zhang, A. Cichocki, Tensor ring decomposition, arXiv preprint arXiv:1606.05535.
- [24] Q. Zhao, M. Sugiyama, A. Cichocki, Learning efficient tensor representations with ring structure networks, arXiv preprint arXiv:1705.08286.
- [25] W. Wang, V. Aggarwal, S. Aeron, Efficient low rank tensor ring completion, *Rn* 1 (r1) (2017) 1.
- [26] L. Yuan, J. Cao, Q. Wu, Q. Zhao, Higher-dimension tensor completion via low-rank tensor ring decomposition, arXiv preprint arXiv:1807.01589.

- [27] L. Yuan, C. Li, D. Mandic, J. Cao, Q. Zhao, Tensor ring decomposition with rank minimization on latent space: An efficient approach for tensor completion, arXiv preprint arXiv:1809.02288.
- [28] Y. Liu, Z. Long, C. Zhu, Image completion using low tensor tree rank and total variation minimization, IEEE Transactions on Multimedia.
- [29] A. Cichocki, N. Lee, I. Oseledets, A.-H. Phan, Q. Zhao, D. P. Mandic, et al., Tensor networks for dimensionality reduction and large-scale optimization: Part 1 low-rank tensor decompositions, Foundations and Trends® in Machine Learning 9 (4-5) (2016) 249–429.
- [30] R. Gemulla, E. Nijkamp, P. J. Haas, Y. Sismanis, Large-scale matrix factorization with distributed stochastic gradient descent, in: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2011, pp. 69–77.
- [31] T. Maehara, K. Hayashi, K.-i. Kawarabayashi, Expected tensor decomposition with stochastic gradient descent., in: AAAI, 2016, pp. 1919–1925.
- [32] Y. Wang, A. Anandkumar, Online and differentially-private tensor decomposition, in: Advances in Neural Information Processing Systems, 2016, pp. 3531–3539.
- [33] S. Wright, J. Nocedal, Numerical optimization, Springer Science 35 (67-68) (1999) 7.
- [34] J. J. Moré, D. J. Thuente, Line search algorithms with guaranteed sufficient decrease, ACM Transactions on Mathematical Software (TOMS) 20 (3) (1994) 286–307.
- [35] D. M. Dunlavy, T. G. Kolda, E. Acar, Poblano v1. 0: A matlab toolbox for gradient-based optimization, Sandia National Laboratories, Albuquerque, NM and Livermore, CA, Tech. Rep. SAND2010-1422.

- [36] D. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.
- [37] S. Ruder, An overview of gradient descent optimization algorithms, arXiv preprint arXiv:1609.04747.
- [38] B. N. Khoromskij, Tensor numerical methods for multidimensional pdes: theoretical analysis and initial applications, *ESAIM: Proceedings and Surveys* 48 (2015) 1–28.
- [39] J. I. Latorre, Image compression and entanglement, arXiv preprint quant-ph/0510031.
- [40] A. Novikov, D. Podoprikin, A. Osokin, D. P. Vetrov, Tensorizing neural networks, in: *Advances in Neural Information Processing Systems*, 2015, pp. 442–450.
- [41] H. Zhang, W. He, L. Zhang, H. Shen, Q. Yuan, Hyperspectral image restoration using low-rank matrix recovery, *IEEE Transactions on Geoscience and Remote Sensing* 52 (8) (2014) 4729–4743.