ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414606

CANONICAL POLYADIC TENSOR DECOMPOSITION WITH LOW-RANK FACTOR MATRICES

Anh-Huy Phan[1], *Petr Tichavský*[2], *Konstantin Sobolev*[1], *Konstantin Sozykin*[1], *Dmitry Ermilov*[1] *and Andrzej Cichocki*[1]

[1] Skolkovo Institute of Science and Technology (Skoltech), Moscow 143026, Russia
[2] The Czech Academy of Sciences, Institute of Information Theory and Automation
Pod vodarenskou vezi 4, 182 00 Prague 8, Czech Republic

ABSTRACT

This paper proposes a constrained canonical polyadic (CP) tensor decomposition method with low-rank factor matrices. In this way, we allow the CP decomposition with high rank while keeping the number of the model parameters small. First, we propose an algorithm to decompose the tensors into factor matrices of given ranks. Second, we propose an algorithm which can determine the ranks of the factor matrices automatically, such that the fitting error is bounded by a user-selected constant. The algorithms are verified on the decomposition of a tensor of the MNIST hand-written image dataset.

Index Terms— CANDECOMP, PARAFAC, low-rank constraint, rank minimization, tensor decomposition

1. INTRODUCTION

The CANDECOMP/PARAFAC or canonical polyadic tensor decomposition (CPD) seeks the best rank-R tensor approximation to a tensor $\mathcal{Y}$ of the size $I \times J \times K$ in the form of

$$\mathcal{Y} \approx \sum_{r=1}^{R} \boldsymbol{a}_r \circ \boldsymbol{b}_r \circ \boldsymbol{c}_r = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!] \,, \tag{1}$$

where "$\circ$" represents the outer product, $\mathbf{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_R]$, $\mathbf{B} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_R]$ and $\mathbf{C} = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_R]$ are factor matrices of the decomposition. A similar model can be extended straightforwardly to higher-order tensors. The CPD has found applications in various fields, including separation of signals in wireless communication systems, independent component analysis, estimation of temporal and spectral patterns in EEG signals, blind identification [1, 2]. Recently CPD has been used to compress Convolutional Neural Networks (CNN) [3–6]. To obtain the CPD, we can minimize the Frobenius norm of the error of the tensor $\mathcal{Y}$ and its estimate, e.g., using the Alternating Least Squares (ALS) algorithm to update sequentially the factor matrices [7–10], or the non-linear conjugate gradient method [11, 12], the Levenberg-Marquardt (LM) algorithm [13, 14], the Krylov LM algorithm [15, 16], the non-linear least squares (NLS) algorithm [17] to update all the parameters at a time.

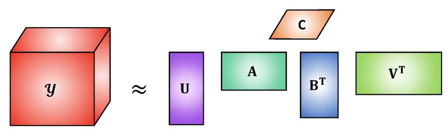


Fig. 1. Illustration of the CPD with low-rank constraints.

In practice, real-world data does not exactly admit the low-rank CPD. Good approximation often requires relatively high rank. For example, to preserve the accuracy of the original CNNs, e.g., ResNets, Alexnet, convolutional kernels can be approximated with ranks of 800, which significantly exceed the tensor dimensions. Such decomposition yields "fat" factor matrices, i.e., with more columns than rows, resulting in high redundancy.

Tucker decomposition (TKD) is another widely used decomposition [18, 19]. This decomposition is particularly suited to compress the tensors before applying CPD but is only applicable when the rank of CPD does not exceed the tensor dimensions. The TKD can be used in the compression of deep neural networks [5], while a combination of TKD+CPD in this application was proposed in [6].

In this paper, we consider the CPD with high rank and propose a new model that seeks a compact subspace to represent the "fat" factor matrices. More specifically, some or all factor matrices are modelled as products of an orthogonal matrix, $\mathbf{U}$ of size $I \times R_1$, $R_1 < I$, and a smaller factor matrix $\mathbf{A}$ of size $R_1 \times R$, R is often much higher than I, $R \gg I$. This results a low-rank constrained CPD (LrCPD), illustrated in Fig. 1

$$\mathcal{Y} \approx [\![\mathbf{UA}, \mathbf{VB}, \mathbf{WC}]\!] \tag{2}$$

where $\mathbf{U}^T\mathbf{U} = \mathbf{I}_{R_1}$, $\mathbf{V}^T\mathbf{V} = \mathbf{I}_{R_2}$ and $\mathbf{W}^T\mathbf{W} = \mathbf{I}_{R_3}$. When $R \leq I$, we can apply the two-stages decomposition, TKD+CPD, which first performs the TKD with multilinear rank $[R_1, R_2, R_3]$, then decomposes the core tensor to get a rank-R tensor [20]. When $R \gg I$, the subspaces sought by TKD may not be optimal, and CPD of the compressed core tensor does not give the best result of the model in (2).

When $\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$ are design matrices, i.e., fixed and known in advance, the model in (2) becomes the CANDELINC [20, 21]. Another related model is when $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are dependence matrices that consist of zeros and ones, we obtain the parallel profiles with linear dependencies (PARALIND) for the rank-overlap problem proposed by R. Bro *et. al.* [22]. A similar constrained PARAFAC (CONFAC) is developed for the blind identification of underdetermined mixtures [23]. A partial uniqueness of PARALIND with doubly linear dependence was investigated in [22, 23]. For applications and variants of PARALIND/CONFAC, we refer to the overview [24]. Note that our proposed constrained CPD can seek the compact subspace to represent factor matrices, while the PARALIND/CONFAC has no constraints on the sub-factor matrices, and the dependence matrices are often given and contain zeros and ones.

For CPD with long (and thin) factor matrices, we can reshape them to higher-order tensors and represent them in the low-rank Tensor Train format [25]. This model is different from ours.

Contribution of the proposed approach includes

- Algorithm for the new constrained CPD with given factor ranks.
- Algorithm for the new model which can search for the best subspace to model the factor matrices in CPD. Ranks of the factor matrices are determined automatically when the approximation error is bounded.

2. CPD WITH LOW-RANK FACTOR MATRICES

We first consider the CPD with ranks of its matrices given, then extend the model to determine the smallest subspace for the factor matrices.

2.1. Factor matrices with given ranks

Assume that the ranks of $\mathbf{A}$ and $\mathbf{B}$ are given, and the model need not factorize the factor matrix $\mathbf{C}$. This case can happen when tensor dimensions are not well balanced, and one dimension can be much smaller than the other ones, e.g., the color layer (3) in the color images, or the filter sizes $(3, 5, 9)$ in the convolutional kernels, or the number of channels in EEG signals compared to the time instants or frequency bins.

Optimization problem for the constrained rank-R CP decomposition can be formulated as

$$\min_{\mathbf{U},\mathbf{V},\mathbf{B},\mathbf{C}} \quad f = \|\mathcal{Y} - [\![\mathbf{UA}, \mathbf{VB}, \mathbf{C}]\!]\|_F^2 \qquad (3)$$
$$\text{s.t.} \quad \mathbf{U}^T\mathbf{U} = \mathbf{I}_{R_1}, \ \mathbf{V}^T\mathbf{V} = \mathbf{I}_{R_2}$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices of size $I \times R_1$ and $J \times R_2$, respectively, $R_1 < I_1$, $R_2 < I_2$ and $R > \max(R_1, R_2)$. The matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ comprise R columns,

We first derive update rules for factor matrices in the first mode. In order to update $\mathbf{A}$, we rewrite the objective function in form of mode-1 unfolding of $\mathcal{Y}$, that is

$$\begin{aligned}
f &= \|\mathcal{Y} - [\![\mathbf{UA}, \mathbf{VB}, \mathbf{C}]\!]\|_F^2 \\
&= \|\mathbf{Y}_{(1)} - \mathbf{UA}(\mathbf{C} \odot \mathbf{VB})^T\|_F^2 \\
&= \|\mathbf{Y}_{(1)}\|_F^2 + \|\mathbf{UA}(\mathbf{C} \odot \mathbf{VB})^T\|_F^2 - 2\,\mathrm{tr}(\mathbf{UA}(\mathbf{C} \odot \mathbf{VB})^T\mathbf{Y}_{(1)}^T) \\
&= \|\mathcal{Y}\|_F^2 + \|\mathbf{A}(\mathbf{C} \odot \mathbf{VB})^T\|_F^2 - 2\,\mathrm{tr}(\mathbf{A}(\mathbf{C} \odot \mathbf{VB})^T(\mathbf{U}^T\mathbf{Y}_{(1)})^T) \\
&= \|\mathcal{Y}\|_F^2 - \|\mathbf{U}^T\mathbf{Y}_{(1)}\|_F^2 + \|\mathbf{U}^T\mathbf{Y}_{(1)} - \mathbf{A}(\mathbf{C} \odot \mathbf{VB})^T\|_F^2,
\end{aligned}$$

where "$\odot$" denotes the Khatri-Rao product. The last expression is achieved due to the orthogonality of $\mathbf{U}$, $\mathbf{U}^T\mathbf{U} = \mathbf{I}_{R_1}$. Let $\mathcal{G} = \mathcal{Y} \times_1 \mathbf{U}^T \times_2 \mathbf{V}^T$. $\mathbf{G}_{(1)}$ is mode-1 unfolding of $\mathcal{G}$. While keeping the other factor matrices fixed, $f()$ achieves its minimum at the optimal $\mathbf{A}^\star$ given by

$$\begin{aligned}
\mathbf{A}^\star &= \mathbf{U}^T\mathbf{Y}_{(1)}(\mathbf{C} \odot \mathbf{VB})^\dagger \\
&= \mathbf{U}^T\mathbf{Y}_{(1)}(\mathbf{C} \odot \mathbf{VB})((\mathbf{C}^T\mathbf{C}) \circledast (\mathbf{B}^T\mathbf{B}))^{-1} \\
&= \mathbf{G}_{(1)}(\mathbf{C} \odot \mathbf{B})((\mathbf{C}^T\mathbf{C}) \circledast (\mathbf{B}^T\mathbf{B}))^{-1}\,, \qquad (4)
\end{aligned}$$

provided that $((\mathbf{C}^T\mathbf{C}) \circledast (\mathbf{B}^T\mathbf{B}))$ is invertible. The symbol "$\circledast$" represents the element-wise Hadamard product. The last expression exploits the following identity

$$\mathbf{C} \odot \mathbf{VB} = (\mathbf{I} \otimes \mathbf{V})(\mathbf{C} \odot \mathbf{B}).$$

Substitute $\mathbf{A}^\star$ into the objective function $f()$, we obtain

$$\begin{aligned}
f(\mathbf{A}^\star) &= \|\mathcal{Y}\|_F^2 - \|\mathbf{U}^T\mathbf{Y}_{(1)}\|_F^2 + \|\mathbf{U}^T\mathbf{Y}_{(1)} - \mathbf{A}^\star(\mathbf{C} \odot \mathbf{VB})^T\|_F^2 \\
&= \|\mathcal{Y}\|_F^2 + \mathrm{tr}(\mathbf{A}^\star(\mathbf{C} \odot \mathbf{VB})^T(\mathbf{C} \odot \mathbf{VB})\mathbf{A}^\star) \\
&\quad -2\,\mathrm{tr}(\mathbf{A}^\star(\mathbf{C} \odot \mathbf{VB})^T\mathbf{Y}_{(1)}^T\mathbf{U}) \\
&= \|\mathcal{Y}\|_F^2 - \mathrm{tr}(\mathbf{U}^T\,\mathbf{Q}\,\mathbf{U})
\end{aligned}$$

where

$$\mathbf{Q} = \mathbf{Y}_{(1)}(\mathbf{C} \odot \mathbf{VB})^T((\mathbf{C}^T\mathbf{C}) \circledast (\mathbf{B}^T\mathbf{B}))^{-1}(\mathbf{C} \odot \mathbf{VB})\mathbf{Y}_{(1)}^T\,. \qquad (5)$$

The last expression of $f()$ in the above equation implies that the optimal $\mathbf{U}$ which minimizes f is the solution of a trace maximization problem

$$\max \quad \mathrm{tr}(\mathbf{U}^T\mathbf{Q}\mathbf{U}) \quad \text{s.t.} \quad \mathbf{U}^T\mathbf{U} = \mathbf{I}_{R_1} \qquad (6)$$

that is, $\mathbf{U}$ comprises R_1 principle eigenvectors of $\mathbf{Q}$. After updating $\mathbf{U}$, the algorithm updates $\mathbf{A}$ using (4).

The factor matrices, $\mathbf{V}$ and $\mathbf{B}$, can be updated similarly. The factor matrix without constraint, e.g., $\mathbf{C}$, is updated using the ordinary ALS update rule, i.e.,

$$\begin{aligned}
\mathbf{C} &= \mathbf{Y}_{(3)}(\mathbf{VB} \odot \mathbf{UA})(((\mathbf{UA})^T(\mathbf{UA}) \circledast ((\mathbf{VB})^T(\mathbf{VB})))^{-1} \\
&= \mathbf{Y}_{(3)}(\mathbf{V} \otimes \mathbf{U})(\mathbf{B} \odot \mathbf{A})((\mathbf{A}^T\mathbf{A}) \circledast (\mathbf{B}^T\mathbf{B}))^{-1} \\
&= \mathbf{G}_{(3)}(\mathbf{B} \odot \mathbf{A})((\mathbf{A}^T\mathbf{A}) \circledast (\mathbf{B}^T\mathbf{B}))^{-1}. \qquad (7)
\end{aligned}$$

In summary, the update rules for $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are similar to the ALS updates for CPD of the tensor $\mathcal{G}$ of size $R_1 \times R_2 \times I_3$, i.e., have lower complexity than those in the ordinary CPD. However, the updates for $\mathbf{U}$ and $\mathbf{V}$ demand an extra cost for the EVD of the matrix $\mathbf{Q}$.

2.2. Optimal Ranks for Factor Matrices

The ranks of the factor matrices may not always be given. In this subsection, we derive a more compact model with the smallest number of parameters so that the rank of factor matrices $\mathbf{UA}$, $\mathbf{VB}$ can be determined automatically based on an approximation error bound. More specifically, for a CPD with given rank R, we find a constrained decomposition, $\hat{\mathcal{Y}} = [\![\mathbf{UA}, \mathbf{VB}, \mathbf{C}]\!]$ with the smallest number of parameters, e.g., we minimize

$$\begin{array}{ll} \min & (I+R)R_1 + (J+R_2)R_2 + KR \\ \text{s.t.} & \|\mathcal{Y} - [\![\mathbf{UA}, \mathbf{VB}, \mathbf{C}]\!]\|_F^2 \leq \varepsilon \|\mathcal{Y}\|_F^2 \end{array}$$

where $0 \leq \varepsilon < 1$.

Define $\mathbf{F} = \mathbf{UA}$. Then $\text{rank}(\mathbf{F}) = R_1$ and $\mathbf{F}$ is solution of the following rank minimization problem

$$\begin{array}{ll} \min & \text{rank}(\mathbf{F}) \\ \text{s.t.} & \|\mathbf{Y}_{(1)} - \mathbf{F}(\mathbf{C} \odot \mathbf{VB})^T\|_F^2 \leq \varepsilon \|\mathcal{Y}\|_F^2. \end{array}$$

Next, we assume that $\mathbf{C} \odot \mathbf{VB}$ is of full column rank. Denote SVD of $\mathbf{C} \odot \mathbf{VB} = \mathbf{ZDK}^T$, where $\mathbf{Z}$ is of size $I_2 I_3 \times R$, $\mathbf{K}$ of size $R \times R$. By exploiting the identity in (5) and orthogonality of $\mathbf{V}$, the SVD of $\mathbf{C} \odot \mathbf{VB}$ is computed through SVD of a smaller matrix $(\mathbf{C} \odot \mathbf{B})$.

The constraint in the above rank minimization problem can be rewritten as

$$\begin{array}{rcl} \|\mathbf{Y}_{(1)} - \mathbf{F}(\mathbf{C} \odot \mathbf{VB})^T\|_F^2 & = & \|\mathcal{Y}\|_F^2 - \|\mathbf{Y}_1 \mathbf{Z}\|_F^2 + \|\mathbf{Y}_1 \mathbf{Z} - \mathbf{FKD}\|_F^2 \\ & \leq & \varepsilon \|\mathcal{Y}\|_F^2 \end{array}$$

Since $\text{rank}(\mathbf{F}) = \text{rank}(\mathbf{FKD})$, we can formulate an equivalent rank-minimization problem for $\tilde{\mathbf{F}} = \mathbf{FKD}$

$$\begin{array}{ll} \min & \text{rank}(\tilde{\mathbf{F}}) \\ \text{s.t.} & \|\mathbf{Y}_{(1)}\mathbf{Z} - \tilde{\mathbf{F}}\|_F^2 \leq \|\mathbf{Y}_{(1)}\mathbf{Z}\|_F^2 - (1-\varepsilon)\|\mathcal{Y}\|_F^2. \end{array}$$

It is straightforward to see that the optimal $\tilde{\mathbf{F}}$ is achieved through the truncated SVD of the matrix, $\mathbf{Y}_{(1)}\mathbf{Z}$, which obeys the bound $\|\mathbf{Y}_{(1)}\mathbf{Z} - \tilde{\mathbf{F}}\|_F^2 \leq \|\mathbf{Y}_1\mathbf{Z}\|_F^2 - (1-\varepsilon)\|\mathcal{Y}\|_F^2$

$$\tilde{\mathbf{F}} = \mathbf{US}\tilde{\mathbf{V}}^T \approx \mathbf{Y}_{(1)}\mathbf{Z}$$

where $\mathbf{S} = \text{diag}(\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{R_1})$ and the rank R_1 is determined as the smallest number of singular values such that $\sum_{r=1}^{R_1} \sigma_r^2 \geq (1-\varepsilon)\|\mathcal{Y}\|_F^2 > \sum_{r=1}^{R_1-1} \sigma_r^2$. In other words, $\mathbf{U}$ comprises the left leading singular vectors of $\mathbf{Y}_{(1)}\mathbf{Z}$ and

$$\mathbf{A} = \mathbf{S}\tilde{\mathbf{V}}^T \mathbf{D}^{-1} \mathbf{K}^T \tag{8}$$

Similar update rule can be derived for $\mathbf{V}$ and $\mathbf{B}$.

Remark 1. *Factor matrices in the low-rank constrained CPD can be initialized by full factor matrices in the ordinary CPD, i.e.,* $\mathbf{U}$ *(and* $\mathbf{V}$*) can first start with an identity matrix of size* $\mathbf{I}_I$*. The rank of the factor matrix* $\mathbf{UA}$ *will be gradually updated to achieve the smallest model.*

Algorithm 1: Low-rank constrained CPD

Input: $\mathcal{Y}$, CP-rank R and an error bound ε
Output: $\hat{\mathcal{Y}} = [\![\mathbf{U}_1\mathbf{A}_1, \mathbf{U}_2\mathbf{A}_2, \ldots, \mathbf{U}_N\mathbf{A}_N]\!]$, $\mathbf{U}_n^T\mathbf{U}_n = \mathbf{I}_{R_n}$ s.t. $\|\mathcal{Y} - \hat{\mathcal{Y}}\|_F^2 \leq \varepsilon\|\mathcal{Y}\|_F^2$
begin
1 Initialize a rank-R CPD of $\mathcal{Y}$
2 Initialize $\mathbf{U}_n = \mathbf{I}_{I_n}$, $n = 1, \ldots, N$
for $n = 1, \ldots, N$ **do**
3 SVD of the Khatri-Rao product $\mathbf{A}_N \odot \cdots \odot \mathbf{A}_{n+1} \odot \mathbf{A}_{n-1} \circ \cdots \odot \mathbf{A}_1 = \mathbf{ZDK}$
4 $\mathcal{T} = \mathcal{Y} \times_{k \neq n} \mathbf{U}_k$
5 **Update** $\mathbf{U}_n$ from truncated-SVD of $\mathbf{T}_{(n)}\mathbf{Z} \approx \mathbf{U}_n\mathbf{S}\mathbf{V}^T$ s.t. $\sum_{r=1}^{R_n} \sigma_r^2 \geq (1-\varepsilon)\|\mathcal{Y}\|_F^2$
6 **Update** $\mathbf{A}_n = \mathbf{S}\mathbf{V}^T\mathbf{D}^{-1}\mathbf{K}^T$

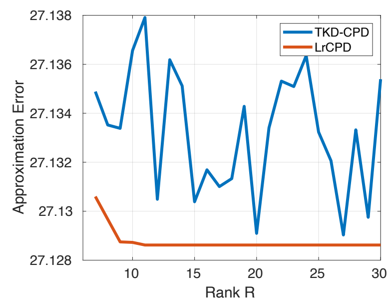


Fig. 2. Decomposition of the Pepper image using LrCPD and TKD+CPD.

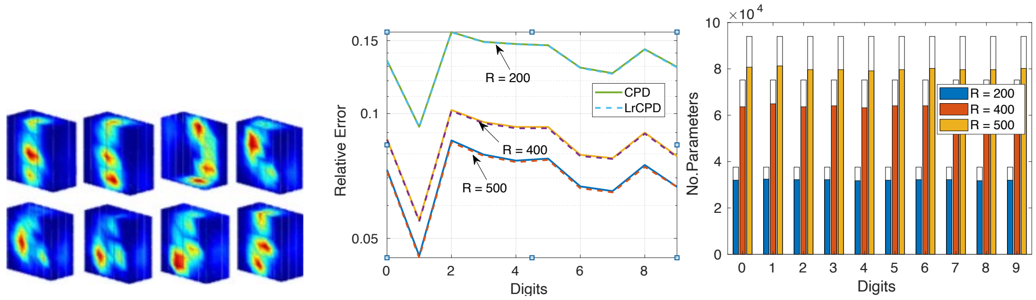
Remark 2. *The error bound,* $\varepsilon\|\mathcal{Y}\|_F^2$*, to control the complexity of the model, can be set to the noise level or the approximation error of the CPD without constraints.*

Remark 3. *For the task to reduce the number of parameters in CPD, the rank* $R_1 < \frac{RI}{R+I}$ *in order to keep the total number of parameters of* $\mathbf{U}$ *and* $\mathbf{A}$ *is smaller than that of the full factor matrix* $\mathbf{F} = \mathbf{UA}$*.*

The proposed algorithm can be similarly extended to higher-order tensors. We skip the detailed derivation and summarize the pseudo-code of the algorithm in Algorithm 1. Step-3 computes Khatri-Rao product of all-but-one subfactor matrices $\mathbf{A}_k$, $k \neq n$.

3. EXPERIMENTAL RESULTS

Example 1 The aim of this example is to compare the proposed algorithm to the naive method which first performs TKD, then CPD of the core tensor of the estimated tensor. We decompose the Pepper image of size $128 \times 128 \times 3$ into three factor matrices with fixed ranks $R_1 = R_2 = 6$ and $R_3 = 2$,



(a) (b) (c)

Fig. 3. (a) Illustration of a Gabor tensor of a handwritten digit image which comprises 8 sub-tensors of Gabor coefficients in 8 orientations and 4 scale layers, (b)Relative errors in approximation of the Gabor tensors with rank $R = 500$ using CPD and LrCPD, (c) Number of parameters in the CPD (white (blank) boxes) and the low-rank constrained CPD shading boxes.

but different number of columns $R = 7, 8, \ldots$. The parameters in CPD are initialized randomly. Approximation errors in Fig. 2 show that LrCPD gives a better approximation than TKD+CPD. Moreover, CPD of the core tensor is sensitive to the initial values and unstable because the decomposition is with rank exceeding the tensor dimensions. Its approximation error does not always decrease with increasing the rank. The fixed subspaces obtained by TKD are not optimal for the low-rank constrained CPD. Similar results have been observed for other values of the ranks $R_1 = R_2$.

Example 2 We demonstrate the proposed model's performance, especially the algorithm for the model with factor ranks optimally determined in Section 2.2. We use the handwritten digit images in the MNIST dataset. From 100 images for each digit, which are of size 28×28, we computed their Gabor features with 8 orientations and 4 scales to yield order-4 tensors of size $28 \times 28 \times 32 \times 100$. The Gabor features of MNIST handwritten digits can be used to cluster or classify the digits [26]. A Gabor tensor with 8 orientations of a digit image is illustrated in Fig 3(a). The CPD and the constrained CPD with optimal factor ranks decomposed the tensors with the same ranks $R = 200, 400$ and 500. Note that for the decomposition with rank exceeding the dimension, TKD cannot be used as a compression tool prior to CPD.

For CPD, we ran the ALS algorithm with line search [9] within 1000 iterations. The obtained tensors were used to initialize the constrained CPD with optimal ranks for factor matrices. We seek low-rank models for all four-factor matrices. The approximation error bound for the low-rank constrained CPD is set to the approximation error attained by CPD, i.e.

$$\varepsilon = \frac{\|\mathcal{Y} - \mathcal{Y}_{cpd}\|_F^2}{\|\mathcal{Y}\|_F^2} .$$

Because the approximation error in the constrained CPD is bounded, the newly yielded models will not be worse than those by CPD. The approximation errors plotted in Fig. 3(b) even show that we obtained lower approximation errors. More importantly, the new models have a fewer number of parameters than the CPD results.

Comparison for the decomposition of tensors for each digit is provided in Fig. 3(c). Blank bars indicate the number of parameters in the CPD of a Gabor tensor. Shading bars show the number of parameters in the constrained CPD. For the same CP-ranks $R = 200, 400, 500$, the constrained CPD yields models with less 5547, 11376, and 14002 parameters on average for all digits images than the ordinary CPD, respectively. As an example, the Gabor tensor for the digit-9 was decomposed with ranks $R_1 = 15$, $R_2 = 13$ and $R_3 = 29$. The rank $R_4 = 89$ for the last mode exceeds $84 (= [500 \times 100/600])$, hence the last factor matrix was not factorized.

4. CONCLUSIONS

The proposed method is a novel method of tensor compression that may be suitable in cases when the tensor itself does not have a low rank. The number of the parameters can still be reduced, as we show with MNIST handwritten image database. Our recent study shows that the combination of Tucker and CPD can provide lower compression ratios for compression of ResNet18 while maintaining the accuracy of the original CNNs [6]. The proposed constrained CPD works similarly to the TKD-CPD decomposition and can be applied to CNN compression.

Acknowledgements

The work of A.-H. Phan, K. Sobolev, K. Sozykin, D. Ermilov and A. Cichocki was supported by the Ministry of Education and Science of the Russian Federation under Grant 14.756.31.0001. The computing for this project was performed on the Zhores CDISE HPC cluster at Skoltech [27].

5. REFERENCES

[1] T.G. Kolda and B.W. Bader, "Tensor decompositions and applications," Tech. Rep. SAND2007-6702, Sandia National Laboratories, Albuquerque, NM and Livermore, CA, November 2007.

[2] A. Cichocki, D. P. Mandic, A.-H. Phan, C. Caifa, G. Zhou, Q. Zhao, and L. De Lathauwer, "Tensor decompositions for signal processing applications. from two-way to multiway component analysis," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 145–163, 2015.

[3] V. Lebedev, Y. Ganin, M. Rakhuba, I. Oseledets, and V. Lempitsky, "Speeding-up convolutional neural networks using fine-tuned CP-decomposition," *International Conference on Learning Representations*, 2015.

[4] X. Zhang, J. Zou, K. He, and J. Sun, "Accelerating very deep convolutional networks for classification and detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, pp. 1943–1955, 2016.

[5] Y.-D. Kim, E. Park, S. Yoo, T. Choi, L. Yang, and D. Shin, "Compression of deep convolutional neural networks for fast and low power mobile applications," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[6] A.-H. Phan, K. Sobolev, K. Sozykin, D. Ermilov, J. Gusak, P. Tichavský, V. Glukhov, I. Oseledets, and A. Cichocki, "Stable low-rank tensor decomposition for compression of convolutional neural network," in *Computer Vision – ECCV 2020*, Cham, 2020, pp. 522–539, Springer International Publishing.

[7] R.A. Harshman, "Foundations of the PARAFAC procedure: Models and conditions for an explanatory multimodal factor analysis," *UCLA Working Papers in Phonetics*, vol. 16, pp. 1–84, 1970.

[8] P. Comon, X. Luciani, and A. L. F. de Almeida, "Tensor decompositions, alternating least squares and other tales," *Journal of Chemometrics*, vol. 23, 2009.

[9] A.-H. Phan, P. Tichavský, and A. Cichocki, "Fast alternating LS algorithms for high order CANDECOMP/PARAFAC tensor factorizations," *Signal Processing, IEEE Transactions on*, vol. 61, no. 19, pp. 4834–4846, 2013.

[10] P. Tichavský, A.-H. Phan, and A. Cichocki, "Partitioned alternating least squares technique for canonical polyadic tensor decomposition," *IEEE Signal Processing Letters*, vol. 23, no. 7, pp. 993–997, July 2016.

[11] P. Paatero and U. Tapper, "Positive matrix factorization: A nonnegative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, pp. 111–126, 1994.

[12] E. Acar, D. M. Dunlavy, and T. G. Kolda, "A scalable optimization approach for fitting canonical tensor decompositions," *Journal of Chemometrics*, vol. 25, no. 2, pp. 67–86, February 2011.

[13] P. Paatero, "A weighted non-negative least squares algorithm for three-way PARAFAC factor analysis," *Chemometrics Intelligent Laboratory Systems*, vol. 38, no. 2, pp. 223–242, 1997.

[14] A.-H. Phan, P. Tichavský, and A. Cichocki, "Low complexity damped Gauss-Newton algorithms for CANDECOMP/PARAFAC," *SIAM Journal on Matrix Analysis and Applications*, vol. 34, no. 1, pp. 126–147, 2013.

[15] P. Tichavský, A.-H. Phan, and A. Cichocki, "Sensitivity in tensor decomposition," *IEEE Signal Processing Letters*, vol. 26, no. 11, pp. 1653–1657, 2019.

[16] P. Tichavský, A. Phan, and A. Cichocki, "Weighted Krylov-Levenberg-Marquardt method for canonical polyadic tensor decomposition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 3917–3921.

[17] L. Sorber, M. Van Barel, and L. De Lathauwer, "Optimization-based algorithms for tensor decompositions: Canonical polyadic decomposition, decomposition in rank-$(L_r, L_r, 1)$ erms, and a new generalization," *SIAM Journal on Optimization*, vol. 23, no. 2, pp. 695–720, 2013.

[18] L. R. Tucker, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, vol. 31, no. 3, pp. 279–311, 1966.

[19] L. De Lathauwer, B. De Moor, and J. Vandewalle, "On the best rank-1 and rank-(R1,R2,...,RN) approximation of higher-order tensors," *SIAM Journal of Matrix Analysis and Applications*, vol. 21, no. 4, pp. 1324–1342, 2000.

[20] R. Bro and C.A. Andersson, "Improving the speed of multiway algorithms - Part II: Compression," *Chemometrics and Intelligent Laboratory Systems*, vol. 42, pp. 105–113, 1998.

[21] J.D. Carroll, S. Pruzansky, and J.B. Kruskal, "Candelinc: A general approach to multidimensional analysis of many-way arrays with linear constraints on parameters," *Psychometrika*, vol. 45, no. 1, pp. 3–24, 1970.

[22] R. Bro, R. A. Harshman, N. D. Sidiropoulos, and M. E. Lundy, "Modeling multi-way data with linearly dependent loadings," *Journal of Chemometrics*, vol. 23, no. 7-8, pp. 324–340, 2009.

[23] A. L. F. de Almeida, X. Luciani, A. Stegeman, and P. Comon, "Confac decomposition approach to blind identification of underdetermined mixtures based on generating function derivatives," *IEEE Transactions on Signal Processing*, vol. 60, no. 11, pp. 5698–5713, 2012.

[24] G. Favier and A. L. F. de Almeida, "Overview of constrained PARAFAC models," *EURASIP J. Adv. Signal Process.*, vol. 2014, pp. 142, 2014.

[25] I. V. Oseledets, "Tensor-train decomposition," *SIAM J. Sci. Comput.*, vol. 33, no. 5, pp. 2295–2317, Sept. 2011.

[26] S. Luan, B. Zhang, S. Zhou, C. Chen, J. Han, W. Yang, and J. Liu, "Gabor convolutional networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, pp. 1254–1262.

[27] I. Zacharov, R. Arslanov, M. Gunin, D. Stefonishin, A. Bykov, S. Pavlov, O. Panarin, A. Maliutin, S. Rykovanov, and M. Fedorov, "Zhores — petaflops supercomputer for data-driven modeling, machine learning and artificial intelligence installed in Skolkovo Institute of Science and Technology," *Open Engineering*, vol. 9, no. 1, 2019.