Tensor Train Decomposition for Efficient Memory Saving in Perceptual Feature-maps

Taehyeon Kim Jieun Lee Yoonsik Choe
Department of Electrical and Electronic Engineering, Yonsei University
{pyomu, jieun_lee, yschoe}@yonsei.ac.kr

Abstract

The perceptual loss functions have been used successfully in image transformation for capturing high-level features from images in pre-trained convolutional neural networks (CNNs). Standard perceptual losses require numerous parameters to compare differences in feature-maps on both an input image and a target image; thus, it is not affordable for resource-constrained devices in terms of utilizing a feature-maps. Hence, we propose a compressed perceptual losses oriented Tensor Train (TT) decomposition on the feature-maps. Additionally, to decide an optimal TT-ranks, the proposed algorithm used the global analytic solution of Variational Bayesian Matrix Factorization (VBMF). Therefore, in proposed method, the low-rank approximated feature-maps consist of salient features by virtue of these two techniques. To the best of our knowledge, we are the first to consider curtailing redundancies in feature-maps via low-rank TT-decomposition. Experimental results in style transfer tasks demonstrate that our method not only yields similar qualitative and quantitative results as that of the original version, but also reduces memory requirement by approximately 77%.

1. Introduction

Along with computer vision and deep-learning technology, perceptual losses are regarded as an important loss function in image transformation tasks such as, image denoising [21], super resolution [11, 19], image-to-image translation [5], etc. Perceptual losses can capture perceptual differences between an input image and a ground-truth image using their feature-maps of pretrained convolutional neural networks (CNNs) [6, 4]. Even though, the perceptual losses are extremely powerful, the large scale of their target image's feature-maps consumes considerable storage and memory bandwidth. In addition, this attribute is the main reason of high computational complexity of perceptual losses. Thus, these disadvantages complicate the implementation of perceptual losses on resource-constrained devices.

To handle these problems, we propose a Tensor Train (TT) decomposition [14] based method to compress perceptual losses. The key contributions of proposed method are:

- A proposed algorithm consists of two simple steps: (1) TT-ranks selection and (2) low-rank TT-decomposition with determined TT-ranks.
- In proposed scheme, TT-decomposition [14] with the rank selected by a global analytic solution of variational Bayesian matrix factorization (VBMF) [13] is applied on a pretrained CNN's feature-maps of target images. Note that we regard the dominant features on ground-truth image's feature-maps, which are obtained through low-rank Tensor-Train decomposition, as salient features.
- Due to the low-rank approximated feature-maps via TT-decomposition can implement basic linear algebraic operations [14], the proposed method does not need to be reconstructed with the original form of tensor to calculate the error value of perceptual losses.
- To evaluate the compressed perceptual losses in both qualitative- and quantitative-way, style transfer experiment based on perceptual losses is conducted [4].

The remainder of the paper is organized as follows. Section 2 reviews preliminaries, Section 3 explains proposed method, Section 4 shows experimental results, and Section 5 summarizes the paper.

2. Preliminaries

Throughout the paper, the N-way tensor is denoted by Euler script letters, *e.g.*, $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$, and matrices are denoted by boldface capital letters, *e.g.*, $\mathbf{A}$. The mode-n unfolding matrix of a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ is denoted by $\mathcal{X}_{(n)} \in \mathbb{R}^{I_1 \cdots I_n \times I_{n+1} \cdots I_N}$. The mode-$n$ product of tensor $\mathcal{X} \in \mathbb{R}^{W \times H \times \cdots \times I_N}$ and matrix $\mathbf{U} \in \mathbb{R}^{J \times I_n}$ is defined by $\mathcal{X} \times_n \mathbf{U}$ and is of size $I_1 \times \cdots \times I_{n-1} \times J \times I_{n+1} \times \cdots \times I_N$.

2.1. Perceptual Losses

A perceptual losses consists of feature reconstruction loss and style reconstruction loss. Both of them measure the high-level perceptual and semantic differences between images; they use feature-maps $\mathcal{F}$ from pretrained CNNs to define the objective function [6, 4]. The feature-maps of the jth layer of the CNNs on image y are represented by $\mathcal{F}^j(y) \in \mathbb{R}^{W \times H \times D}$ where W, H, and D are the width, height, and depth of the feature-maps, respectively.

2.1.1 Feature Reconstruction Loss

The feature reconstruction loss is the Euclidean distance between the feature-maps of the target image y and input image $\hat{y}$:

$$L^{j,\mathcal{F}}_{feat}(y, \hat{y}) = \frac{1}{WHD} ||\mathcal{F}^j(y) - \mathcal{F}^j(\hat{y})||^2_2. \quad (1)$$

Using a feature reconstruction loss can encourage the input image $\hat{y}$ to be perceptually similar to the target image y, but does not force it to match exactly with the target image [6, 4].

2.1.2 Style Reconstruction Loss

The style reconstruction loss is squared Frobenius norm of the difference between the Gram matrices of $\mathcal{F}^j(\hat{y})$ and $\mathcal{F}^j(y)$:

$$L^{j,\mathcal{F}}_{style}(y, \hat{y}) = ||G(\mathcal{F}^j(y)) - G(\mathcal{F}^j(\hat{y}))||^2_F, \quad (2)$$

where $G(\mathcal{F}^j(\cdot))$ is the Gram matrix of $\mathcal{F}^j(\cdot)$ and is of size $D \times D$. The style reconstruction loss preserves the stylistic features of the ground-truth image y, but does not maintain its content information such as spatial structure [6, 4].

2.2. Tensor Decomposition

The tensor decomposition is a form of high-order principal component analysis [9]. It decomposes a tensor into a core tensor multiplied by matrices along each mode. For example, when we consider a three-way tensor case $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$,

$$\mathcal{X} \approx \mathcal{G} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3 \equiv [\![\mathcal{G}; \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3]\!] \quad (3)$$

where $\mathbf{U}_1 \in \mathbb{R}^{I_1 \times R_1}, \mathbf{U}_2 \in \mathbb{R}^{I_2 \times R_2}$ and $\mathbf{U}_3 \in \mathbb{R}^{I_3 \times R_3}$ are factor matrices that can be considered as the principal components in each mode. $\mathcal{G} \in \mathbb{R}^{R_1 \times R_2 \times R_3}$ is called the core tensor and its entries indicate the interaction level between the different components. R_n is rank of the mode-n matricization of the tensor which has a size of $I_n \times I_1 \cdots I_{n-1} I_{n+1} \cdots I_3$. If the core tensor is superdiagonal and $R_1 = R_2 = R_3$, this is called the CP decomposition [7]; otherwise, this is called the Tucker decomposition [18]. Perceptual losses are based on the linear algebraic computation of tensors. However, since these decompositions can not perform linear algebraic operations in their decomposed state, the original size of the tensor must be reconstructed. In order to compress perceptual losses, hence, the proposed method used the TT-decomposition which can perform the linear algebraic operations in decomposed state [14, 2].

Table 1. Storage complexities of tensor decompositions for a N-way tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$. $I = \max\{I_1, ..., I_N\}$ and R is the maximum rank value of each tensor decompositions, *e.g.*, in TT-decomposition $R = \max\{R_0, ..., R_N\}$.

Tensor format	Storage complexity
Full tensor	$\mathcal{O}(I^N)$
CP-decomposed tensor	$\mathcal{O}(NIR)$
Tucker-decomposed tensor	$\mathcal{O}(NIR + R^N)$
TT-decomposed tensor	$\mathcal{O}(NIR^2)$

3. Proposed Method

3.1. Low-rank approximated Feature-maps via TT-decomposition

In order to reduce the storage complexity of perceptual losses, proposed method utilizes TT-decomposition. TT-decomposition decomposes target image's feature-maps, $\mathcal{F}^j(y) \in \mathbb{R}^{W \times H \times D}$, into sparsely interconnected three core tensors. Following consecutive expressions for the low-rank approximated feature-maps $\mathcal{F}^j(y)$ via TT-decomposition are as follows:

$$\mathcal{F}^j_{(1)}(y) = \mathbf{U}^{(1)} \mathbf{V}^{(1)} \in \mathbb{R}^{W \times HD}, \quad (4)$$

$$\mathcal{V}^{(1)} = \text{reshape}(\mathbf{V}^{(1)}) \in \mathbb{R}^{R_1 \times H \times D}, \quad (5)$$

$$\mathcal{G}^{(1)} = \text{reshape}(\mathbf{U}^{(1)}) \in \mathbb{R}^{R_0 \times W \times R_1}, \quad (6)$$

where the $\mathbf{U}^{(1)}$ and $\mathbf{V}^{(1)}$ are factor matrices of sizes $W \times R_1$ and $R_1 \times HD$, respectively. The $\mathcal{G}^{(1)}$ is the first core tensor in TT-format of $\mathcal{F}^j(y)$. Since TT-decomposition imposed the boundary condition, R_0 is equivalent to 1,

$$\mathcal{V}^{(1)}_{(2)} = \mathbf{U}^{(2)} \mathbf{V}^{(2)} \in \mathbb{R}^{R_1 H \times D}, \quad (7)$$

$$\mathcal{G}^{(2)} = \text{reshape}(\mathbf{U}^{(2)}) \in \mathbb{R}^{R_1 \times H \times R_2}, \quad (8)$$

$$\mathcal{G}^{(3)} = \text{reshape}(\mathbf{V}^{(2)}) \in \mathbb{R}^{R_2 \times D \times R_3}, \quad (9)$$

where the $\mathbf{U}^{(2)}$ and $\mathbf{V}^{(2)}$ are factor matrices of sizes $R_1 H \times R_2$ and $R_2 \times D$, respectively. $\mathcal{G}^{(2)}$ and $\mathcal{G}^{(3)}$ is second- third-core tensor in TT-format of $\mathcal{F}^j(y)$. Also due to the boundary condition, R_3 becomes 1. Thus, the rank-(R_0, R_1, R_2, R_3) TT-decomposition of the feature-maps $\mathcal{F}^j(y)$ has the form:

$$\mathcal{F}^j(y) \approx \psi(\mathcal{F}^j(y)) = \mathcal{G}^{(1)} \times_1 \mathcal{G}^{(2)} \times_1 \mathcal{G}^{(3)}, \qquad (10)$$

where $\psi(\mathcal{F}^j(y))$ is low-rank approximated feature-maps via TT-decomposition. $\mathcal{G}^{(1)} \in \mathbb{R}^{R_0 \times W \times R_1}$, $\mathcal{G}^{(2)} \in \mathbb{R}^{R_1 \times H \times R_2}$, and $\mathcal{G}^{(3)} \in \mathbb{R}^{R_2 \times D \times R_3}$ are three-dimensional core tensors. The rank set constituting with (R_0, R_1, R_2, R_3) is TT-ranks which are imposed "boundary conditions" $R_0 = R_3 = 1$.

Table 1 shows storage complexities of several tensor decompositions including CP-decomposition, Tucker-decomposition and TT-decomposition. It can be seen that TT-decomposition has asymptotically the same storage complexity as the CP-decomposition, but since the computation of TT-decomposition is based on low-rank approximation of auxiliary unfolding matrices, it is more stable than CP-decomposition. We regard the $\psi(\mathcal{F}^j(y))$ as the low-rank approximated feature-maps which are consisted with only the most salient features. The main reason for using TT decomposition to compress perceptual losses is that it can perform linear algebraic operations in its decomposed state. In [14], they show how TT-format tensors perform basic linear algebraic operations such as addition, scalar product, matrix-by-vector product, and norms. Therefore, in order to compute the perceptual losses, proposed method has no need to reconstruct the decomposed feature-maps into original shaped feature-maps.

3.2. TT-rank Selection via global analytic solution of VBMF

The TT-ranks are important hyper-parameters which control the storage complexity of compressed perceptual losses. In this paper, the compression rate M is defined to indicate the number of parameters required in compressed version compared to that of the original perceptual losses;

$$M = \frac{R_1 W + R_1 R_2 H + R_2 D}{WHD}, \qquad (11)$$

where the denominator term and numerator term are represented by the number of parameters of $\mathcal{F}^j(y)$ and $\psi(\mathcal{F}^j(y))$, respectively. Since W, H, and D are already fixed parameters, it can be seen that the compression rate is related with the value of TT-ranks, only. In [14], they show how to select the TT-ranks with tolerance of reconstruction error, but it is time-consuming trial-and-error. Therefore, the proposed method utilizes data-driven one-shot TT-ranks selection method via global analytic solution of VBMF [13]. Recently, the global analytic VBMF is promising rank selection method which is based on empirical Bayes [12] with automatic relevance determination prior [17]. In order to claim the feasibility of TT-ranks selection via global analytic solution of VBMF, we present the proposition based on theorem about TT-ranks in [14].

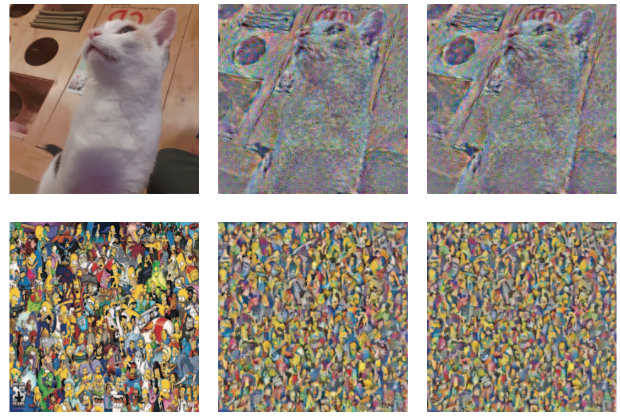

Figure 1. Comparison of visual results for style reconstruction **(top)**; feature reconstruction **(bottom)**. **First column:** examples of target images. **Second column:** the output images via compressed perceptual losses. **Third column:** the output images via standard perceptual losses.

Proposition 1. *If global analytic VBMF can find the rank R_n of unfolding matrix $\mathcal{X}_{(n)}$ with theoretical condition for perfect rank recovery, then the determined TT-ranks via global analytic VBMF can be the upper-bound of TT-ranks, because there exists a TT-decomposition with TT-ranks not higher than R_n.*

Although global analytic VBMF generates suboptimal TT-ranks, it is highly reproducible approach because it makes the proposed method without time consuming TT-rank selection algorithm. Therefore, the proposed method employs the global analytic VBMF in Eq. 4 and 7.

3.3. Compressed Perceptual Losses

The compressed feature reconstruction loss and compressed style reconstruction loss are as follows:

$$\tilde{L}^{j,\mathcal{F}}_{feat}(y,\hat{y}) = \frac{1}{WHD} ||\psi(\mathcal{F}^j(y)) - \mathcal{F}^j(\hat{y})||^2_2, \qquad (12a)$$

$$\tilde{L}^{j,\mathcal{F}}_{style}(y,\hat{y}) = ||G(\psi(\mathcal{F}^j(y))) - G(\mathcal{F}^j(\hat{y}))||^2_F. \qquad (12b)$$

As shown in Fig. 1, the output images from the compressed style reconstruction loss $\hat{y} = \arg\min_{\hat{y}} \tilde{L}^{j,\mathcal{F}}_{style}(y,\hat{y})$ and output images from the compressed feature reconstruction loss $\hat{y} = \arg\min_{\hat{y}} \tilde{L}^{j,\mathcal{F}}_{feat}(y,\hat{y})$ preserve the stylistic features and content information from the target image y, respectively. Moreover, it is difficult to distinguish the differences between the output of original version and that of the compressed version. This implies that we can obtain very similar results from the proposed method qualitatively even though our method requires lower memory than the original loss function.

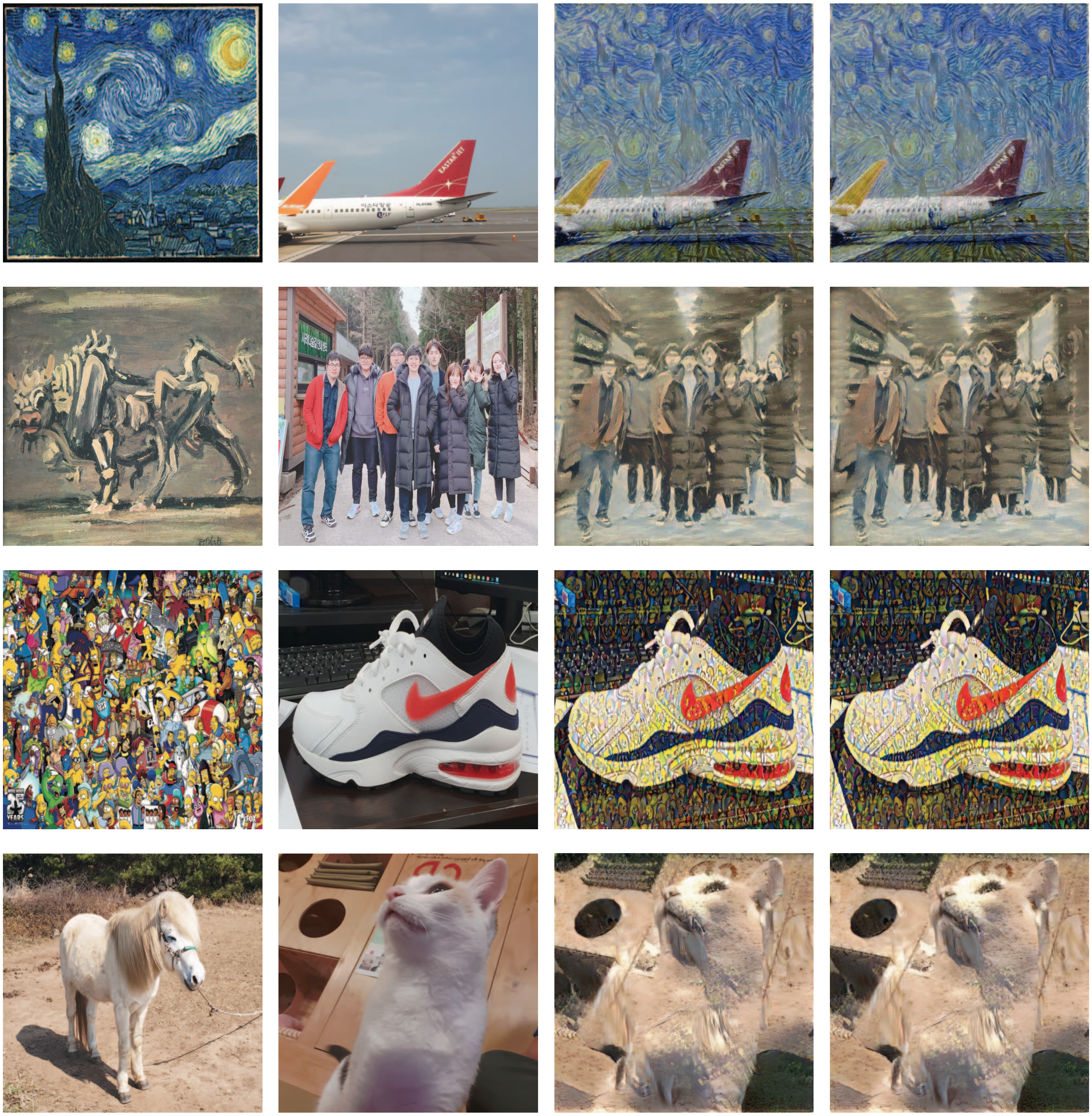

Figure 2. Comparing visual results of style transfer with original perceptual losses [4] and proposed perceptual losses. **First column:** the content images y_c. **Second column:** the style images y_s. **Third column:** the output images of style transfer with *the proposed perceptual losses*. **Fourth column:** the output images of style transfer with the original perceptual losses. In this paper, experiments presented in from **first row** to **fourth row** are referred to as VINCET, LEE, SIMPSONS, and DONKEY, respectively.

4. Experimental Results

To confirm the validity of the proposed method, we designed a style transfer application owing to its simple formulation comprising the superposition between feature reconstruction loss and style reconstruction loss. In addition, we used image quality measures to compare between the output images of the original perceptual losses and those of the compressed perceptual losses [20].

4.1. Style Transfer

The aim of style transfer is to make an input image $\hat{y}$ by inserting target style y_s on the target content image y_c [6, 4]. Thus, we jointly minimize both the compressed feature reconstruction loss and compressed style reconstruction loss. As a baseline, we implemented the primary framework of [4] in our model. Namely, our objective function for style transfer is expressed as follows:

$$\hat{y} = \arg\min_{\hat{y}} \alpha \tilde{L}^{j_1,\mathcal{F}}_{feat}(y_c, \hat{y}) + \beta(\tilde{L}^{j_2,\mathcal{F}}_{style}(y_s, \hat{y}) + \cdots + \tilde{L}^{j_6,\mathcal{F}}_{style}(y_s, \hat{y})), \tag{13}$$

where α and β are the weighting parameters for the feature and style losses, respectively. In addition, $j_1, ..., j_6$ explicitly represent the 10th, 1st, 3rd, 5th, 9th, and 13th layer in a pre-trained VGG-19 [16] on the ImageNet dataset [3], respectively.

4.1.1 Training Details

We used Adam optimizer [8] with a learning rate of 3×10^{-4}. To improve the convergence speed, we set the target image y_c as $\hat{y}$ in the first iteration. The ratio α/β is fixed as 1×10^{-6}. The number of maximum iterations for training is 2×10^3. Our implementation utilizes PyTorch [15], Tensorly [10] and cuDNN [1] on a GTX 1080 8-GB GPU.

4.1.2 Qualitative Results

In Fig. 2, we compare the result images of style transfer from the original perceptual losses and proposed method. Although the proposed method spends much smaller memory than the original method, our experimental results are qualitatively similar to the outcomes from the original losses. For example, both the original method and proposed method tend to preserve objects such as plane, person, shoe, and cat, but remove the content detail on the background information. Therefore, it is verified in Fig. 2 that the proposed algorithm can conserve the performance of the original loss function by low-rank approximated feature-maps which is comprised of dominant features (*i.e.*, TT-decomposition).

4.1.3 Quantitative Results

To analyze the effects of our perceptual losses, we reported the PSNR and SSIM [20] between the output of the proposed method and that of the original version. In Table 2, it can be seen that PSNR and SSIM are more than 25 dB and 90%, respectively, in all experiments. In other words, there is only a small difference between the results of the proposed method and the original method. In addition, the compressed perceptual losses use only almost a quarter of the number of parameters in the original version. According to Fig.2 and Table 2, the proposed method not only has quantitatively and qualitatively close results to the output of original perceptual losses, but also utilizes almost 23% of the memory on average compared to the conventional one.

Table 2. The Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) for the output images via proposed method, with the output images through original method as the reference. The compression rate of proposed method is defined as M.

Experiment	SSIM (%)	PSNR (dB)	M (%)
VINCENT	92.44	28.6488	21.90
LEE	92.53	31.9854	23.88
SIMPSONS	91.90	25.1664	21.37
DONKEY	93.24	30.1937	25.56
Average	92.53	28.9986	23.18

4.2. Limitation and Future work

Since the proposed method can decompose the target image's feature-maps directly, it leads to perform the large scale singular value decomposition which has cubic computational complexities to the number of input size. To handle this difficulties, [14, 2] have proposed the huge scale structured matrices TT-decomposition by transforming the $X \in \mathbb{R}^{I \times J}$ into $\mathcal{X} \in \mathbb{R}^{I_1 \times J_1 \times \cdots I_N \times J_N}$, where $I = I_1 \cdots I_N$ and $J = J_1 \cdots J_N$. Thus, those algorithms could alleviate the large scale matrices decomposition problem present in the proposed method.

5. Conclusion

We herein proposed a novel compressing scheme for perceptual losses based on low-rank TT-decomposition. Further, we extended the feature maps into the low-rank TT-decomposition mechanism to efficiently extract salient features by using only small memory storages. Using two algorithms, *i.e.*, VBMF and TT-decomposition, we obtained the optimal ranks by reducing the required memory with almost 77% reduction and preserved the performance of the original perceptual loss version simultaneously.

Acknowledgement

This work was supported by the Industrial Strategic Technology Development Program (10073229, Development of 4K high-resolution image based LSTM network deep learning process pattern recognition algorithm for real-time parts assembling of industrial robot for manufacturing) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea)

References

[1] S. Chetlur, C. Woolley, P. Vandermersch, J. Cohen, J. Tran, B. Catanzaro, and E. Shelhamer. cuDNN: Efficient primitives for deep learning. *arXiv preprint arXiv:1410.0759*, 2014.

[2] 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):249–429, 2016.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[4] L. A. Gatys, A. S. Ecker, and M. Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[5] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976. IEEE, 2017.

[6] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016.

[7] H. A. Kiers. Towards a standardized notation and terminology in multiway analysis. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 14(3):105–122, 2000.

[8] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[9] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[10] J. Kossaifi, Y. Panagakis, A. Anandkumar, and M. Pantic. Tensorly: Tensor learning in python. *CoRR*, abs/1610.09555, 2018.

[11] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114. IEEE, 2017.

[12] D. J. MacKay. Bayesian interpolation. *Neural computation*, 4(3):415–447, 1992.

[13] S. Nakajima, M. Sugiyama, S. D. Babacan, and R. Tomioka. Global analytic solution of fully-observed variational Bayesian matrix factorization. *Journal of Machine Learning Research*, 14(Jan):1–37, 2013.

[14] I. V. Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317, 2011.

[15] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017.

[16] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[17] M. E. Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1(Jun):211–244, 2001.

[18] L. R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

[19] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018.

[20] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[21] Q. Yang, P. Yan, Y. Zhang, H. Yu, Y. Shi, X. Mou, M. K. Kalra, Y. Zhang, L. Sun, and G. Wang. Low dose CT image denoising using a generative adversarial network with wasserstein distance and perceptual loss. *IEEE transactions on medical imaging*, 2018.