IMAGE DATASET COMPRESSION BASED ON MATRIX PRODUCT STATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale datasets have produced impressive advances in machine learning. However, storing datasets and training neural network models on large datasets have become increasingly expensive. In this paper, we present an effective dataset compression approach based on the matrix product states (MPS) from quantum many-body physics. It can decompose an original image into a sequential product of tensors which effectively retain short-range correlation information in the data for training deep neural networks from scratch. Based on the MPS structure, we propose a new dataset compression method that compresses datasets by filtering long-range correlation information in task-agnostic scenarios and uses dataset distillation to supplement the information in task-specific scenarios. Our approach boosts the model performance by information supplementation, and meanwhile maximizes useful information for the downstream task. Extensive experiments have demonstrated the effectiveness of the proposed approach in dataset compression, especially obtained better model performance (3.19% on average) than state-of-the-art methods for the same compression rate.

1 INTRODUCTION

Large-scale datasets consisting of millions of samples are becoming the norm to obtain state-of-the-art machine learning models in several fields including speech enhancement and recognition (Sun et al., 2020; SainathTN et al., 2013), computer vision (Russakovsky et al., 2015) and natural language processing (Devlin et al., 2019). At such a scale, the resources needed to store datasets and train neural networks become very large, and training machine learning models on it requires the use of specialized equipment and infrastructure. Therefore, it is the critical problems in machine learning that effectively reduce the size of the dataset as well as maintaining the model performance.

An intuitive way is data selection, also known as core-set construction method, *i.e.*, identifying the most representative training samples, which aims to improve the *data efficiency* of machine learning techniques. (Agarwal et al., 2004; Chen, 2009; Feldman et al., 2020), mainly focusing on clustering problems. Another dataset compression method is dataset distillation (Wang et al., 2018; Zhao et al., 2021), which can learn a small set of informative images from large training data and improve the weakness of data selection methods. However, these studies mainly adopt the existing sample reduction techniques to dataset compression, which may not be intrinsically appropriate for architecture of the image dataset. For example, most dataset compression methods need to adopt the whole image information, although only a small proportion of information (*i.e., locality* information) will significantly influence the performance during training. Meanwhile, the performance of models trained the offline compressed datasets on downstream tasks would be restricted.

In this paper, we introduce a novel matrix product states (MPS) technique (Fannes et al., 1992) from quantum many-body physics for compressing image dataset. The MPS is an algorithm that factorizes a matrix into a sequential product of local tensors (*i.e.,* a multi-way array). Here, we call the "locality of pixel dependencies" in image with *short-range correlation* and the "global dependencies" in image with *long-range correlation.* An important merit of the MPS decomposition is that it establishes the structural of *classical pixel correlation* with *quantum entanglement entropy* (Srednicki, 1993) in the dataset: the larger the entanglement entropy is, the less short-range correlation is, and vice versa. The information that has larger classical correlation is mainly short-range correlation in the image dataset (Krizhevsky et al., 2017). The dataset distillation can Such a property motivates us to think about whether such an MPS can be applied to derive a better dataset compression approach. We can compress the dataset by filtering long-range correlation information in task-agnostic scenarios and use the dataset distillation to supplement the information in task-specific scenarios.

To this end, we propose a MPS-based Dataset compression approach, called *MPSD*, to compress the image dataset, which not only enables deep neural networks to obtain similar performance as on the original dataset but also can be used for different models as well as different types of tasks. We have made two import technical contributions for image dataset compression based on MPS. First, we introduce a new task-agnostic dataset compression procedure that efficiently extracting short-range correlation among pixels. We formulate this goal as the problem of minimizing the difference between multiple low-rank tensors with constraints and the original data samples. We present both theoretical discussion and experimental verification for the effectiveness of this dataset compression strategy. Second, we propose a new task-specific module for information supplementation, tailored for machine learning model. Since different downstream tasks have different information for image datasets, the offline dataset compression does not contain task-specific information. We propose a module based on dataset distillation to make the compressed datasets adaptable to different tasks.

To our knowledge, it is the first time that MPS is applied to the image dataset compression, which is well suited for model training and dataset storage. We construct experiments to evaluate the effectiveness of the proposed compression approach for CIFAR, FashinMNSIT, and ImageNet, respectively. Extensive experiments have demonstrated the effectiveness of the proposed approach in dataset compression, especially obtained better model performance (3.19% on average) than similar methods for the same compression rate.

In the rest of the paper, we first review the related work in Section 2. Then we present MPS decomposition and theoretical analysis about quantum entanglement with classical correlation in Section 3. Section 4 introduce our proposed MPS-based dataset compression approach. We report experimental results in Section 5 and conclude the paper in Section 6. We will release code and pre-processed data to reproduce our experiments.

2 Related Work

We review the related works in three aspects.

Core-set Construction The core-set construction technique of selecting the valid knowledge through an illuminating or a priori approach (Toneva et al., 2019; Castro et al., 2018; Aljundi et al., 2019; Sener & Savarese, 2018), either by giving illuminating knowledge about the task or by finding representative samples. The core-set construction define representative criterion in the first (*e.g.,* compactness (Rebuffi et al., 2017; Castro et al., 2018), forgetfulness (Toneva et al., 2019), diversity (Sener & Savarese, 2018; Aljundi et al., 2019)), then select representative samples from original dataset based on the criterion, finally use the selected small dataset to train the machine learning model for a downstream task. In contrast, our approach does not require the presence of a representative sample and is a more general approach.

Knowledge Distillation Knowledge distillation is a technique of transferring knowledge from a collection of models into a single model (Hinton et al., 2015; Buciluǎ et al., 2006; Ba & Caruana, 2014; Romero et al., 2015). While network distillation aims to distill the knowledge of multiple networks into a single model, dataset distillation models network parameters as a function of synthetic training data and learn their synthetic data by minimizing the training loss on the original training data and the synthetic training data (Wang et al., 2018). We use the idea of knowledge distillation to complement the learning of task-relevant information under different tasks. In other words, our goal is to capture the portion of information in the dataset sample that is valid for training deep neural networks, and to perform a "selection" of information in the dataset sample.

Tensor-based Matrix Representation. Tensor-based method of matrices is a technique that allows representing dataset samples in the tensor form such that quantum entanglement corresponds to classical correlations between different coarse-grained textures (Latorre, 2005). Another application is the compression of neural networks. Matrix product operators have been used to compress linear layers of deep neural networks (Gao et al., 2020). The idea of reshaping weights of fully connected layers into high-dimensional tensors and encoding them in Tensor Train format was introduced by Novikov et al. (2015). In contrast, we represent a dataset sample jointly with multiple low-rank tensors, each low-rank tensor describing the difference in information between the previous other tensors and the original graph (*i.e.,* residual information).

Our work is highly built on these studies, while we have a new perspective by designing the dataset compression algorithm which enables extracted short-range correlation in the image. It is the first time that MPS is applied to image dataset compression, and we make contributions for a novel approach to dataset compression.

3 Preliminary

In this paper, scalars are denoted by lowercase letters (*e.g.,* a), matrices are denoted by boldface capital letters (*e.g.,* $\mathbf{M}$), and high-order (order three or higher) tensors are denoted by boldface Euler script letters (*e.g.,* $\mathcal{T}$). A 3-order tensor $\mathcal{T}_{i_1,i_2,i_3}$ can be considered as a (potentially multi-dimensional) array with 3 indices $\{i_1, i_2, i_3\}$.

3.1 Matrix Product State

Originating from quantum many-body physics, matrix product states (MPS) is a standard algorithm to factorize a matrix into a sequential product of multiple local tensors (*i.e.*, a multi-way array) (Latorre, 2005; Perez-Garcia et al., 2007). This MPS decomposition establishes the structure of classical pixel correlation with quantum entanglement entropy. Formally, given a matrix $\mathbf{M} \in \mathbb{R}^{I\times J}$, its MPS decomposition into a product of n local tensors can be represented as:

$$\mathrm{MPS}\ (\mathbf{M}) = \prod_{\mathrm{k=1}}^{\mathrm{n}} \mathcal{T}_{(\mathrm{k})}[\mathrm{d_{k-1}}, \mathrm{j_k}, \mathrm{d_k}], \quad \mathrm{d_k} = \min\left(\prod_{\mathrm{m=1}}^{\mathrm{k}} \mathrm{j_m}, \prod_{\mathrm{m=k+1}}^{\mathrm{n}} \mathrm{j_m}\right), \tag{1}$$

where the $\mathcal{T}_{(k)}[d_{k-1}, j_k, d_k]$ is a 3-order tensor with size $d_{k-1} \times j_k \times d_k$ in which $\prod_{k=1}^{n} j_k = I \times J$ and $d_0 = d_n = 1$. We use the concept of *bond* to connect two adjacent tensors (Fannes et al., 1992). The bond dimension d_k is defined by: we can see from Equation (1) that the d_k is large in the middle and small on both sides. We present a detailed algorithm for MPS decomposition in Algorithm 1. It is usually take an odd number of local tensors with MPS.

Algorithm 1 MPS decomposition for a matrix.

```
Require: matrix $\mathbf{M}$, the number of local tensors $n$
Ensure: : MPS tensor list $\{\mathcal{T}_{(k)}\}_{k=1}^{n}$
for $k = 1 \to n-1$ do
    $\mathbf{M}[I, J] \longrightarrow \mathbf{M}[d_{k-1} \times j_k, -1]$
    $\mathbf{U}\lambda\mathbf{V}^\top = \mathrm{SVD}\ (\mathbf{M})$
    $\mathbf{U}[d_{k-1} \times j_k, d_k] \longrightarrow \mathcal{U}[d_{k-1}, j_k, d_k]$
    $\mathcal{T}^{(k)} := \mathcal{U}$
    $\mathbf{M} := \lambda\mathbf{V}^\top$
end for
$\mathcal{T}^{(n)} := \mathbf{M}$
Normalization
return $\{\mathcal{T}_{(k)}\}_{k=1}^{n}$
```

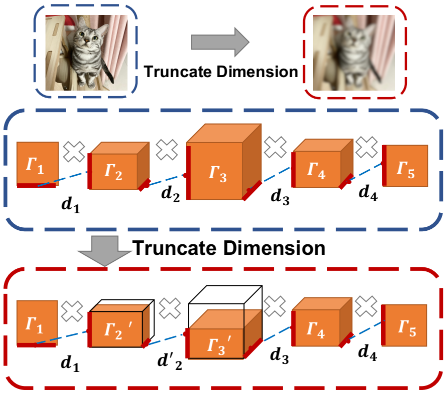

Figure 1: MPS decomposition with five tensors. Dash line denotes *bond* of MPS tensors.

3.2 MPS-based Low-Rank Approximation.

With the MPS decomposition describe in Equation (1), we can exactly decompose a matrix by MPS into the form of a series of products of local tensors and multiply these tensors together to completely reconstruct the original matrix $\mathbf{M}$. We can truncate the k-th bond dimension d_k (see Equation (1)) of local tensors to $d_k^{'}$ for low-rank approximation ($d_k > d_k^{'}$). Different values for $\{d_k\}_{k=1}^{n}$ can be set to control the filtering ability of long-range correlation.

Definition 1. (Local truncation error). Let $\{\lambda_j\}_{j=1}^{d_k}$ are the singular values of $M[j_1, \ldots, j_k, j_{k+1}, \ldots, j_n]$. We define the truncation error induced by the k-th bond dimension d_k local truncation error ϵ_k, which can be efficiently computed as $\epsilon_k = \sum_{j=d_k-d_k^{'}}^{d_k} \lambda_j$.

After defining the local truncation error in Definition 1, we can derive the upper exact bound of the truncation error of the MPS decomposition by iteration.

Theorem 1. *(Truncation error for MPS-based approximation). Let ϵ_k denoted the local truncation error of k-th bond dimension. The upper exact bound of the truncation error with MPS decomposition can be caclulated by:*

$$||\mathbf{M} - \mathrm{MPS}(\mathbf{M})||_F \leq \sqrt{\sum_{k=1}^{n-1} \epsilon_k^2}. \tag{2}$$

The proof can be found in the supplementary materials. Suppose that we have truncated the dimensions of local tensors from $\{d_k\}_{k=1}^{n}$ to $\{d_k^{'}\}_{k=1}^{n}$, the compression ratio can be computed by $\rho = \frac{\sum_{k=1}^{n} d_{k-1}^{'} j_k d_k^{'}}{\prod_{k=1}^{n} j_k}$. The smaller is the compression ratio, the fewer parameters are kept in MPS representation.

3.3 Quantum Entanglement and Classical Correlations

Since the MPS representation can be constructed by operating a series of successive Schmidt decompositions (Vidal, 2004). The entanglement entropy is suitable as a metric to measure correlation information contained in bonds of MPS, which is analogous to entropy in information theory but replaces probabilities with normalized singular values created by SVD. Following (Calabrese & Cardy, 2004), the entanglement entropy with the k-th bond can be calculated by:

$$E_k = -\sum_{j=1}^{d_k} v_j \ln v_j, \quad k = 1, 2, \ldots, n-1 \tag{3}$$

where $\{v_j\}_{j=1}^{d_k}$ denote the normalized SVD eigenvalues of orignal matrix $\mathbf{M}[j_1 \ldots j_k, j_{k+1} \ldots j_n]$. From Equation (3), we note that the larger the entanglement entropy is, the less short-range correlation is, and vice versa.

4 Approach

The short-range correlation information of image datasets is very important for training step. MPS decomposition can effectively filter the short-range correlation information from the image dataset. Hence, it would be natural to apply an MPS-based approximation to compress the image matrices in the dataset by truncating the bond dimensions of MPS. In particular, we propose two main improvements for MPS-based dataset compression, which can efficiently compress the image dataset and effectively complementary task-specific information.

4.1 Task-agnostic Dataset Compression

Suppose we are given a large dataset consisting of $|\mathcal{S}|$ training samples $\mathcal{S} = \{(\mathbf{S}_i)\}_{i=1}^{|\mathcal{S}|}$ where $\mathbf{S}_i \in \mathcal{S} \subset \mathbb{R}^d$, $\mathcal{S}$ is a d-dimensional input space. We denotes the $\mathrm{MPS}(\mathbf{S}_\mathrm{i})$ as the truncated tensor set with MPS decomposition on $\mathbf{S}_i$. Similar with *Image Compression with Entanglement* (ICE), which was proposed to use MPS for truncated compression after performing a discrete cosine Fourier transform of $\mathbf{S}_i$ (Latorre, 2005). We use the MPS decomposition of images to filter long-range correlation information to achieve dataset compression. However, the dataset after this method for compression has a significant information loss since the truncated dimension $\{d_k^{'}\}_{k=1}^{n}$ decreases (this is discussed in Section 3.2.).

To address this problem, inspired by He et al. (2016), we propose to insert residual information (*i.e.,* the difference between $\mathbf{S}_i$ and $\mathrm{MPS}(\mathbf{S}_\mathrm{i})$) to the MPS representation, which is defined by:

$$\mathrm{R}(\mathbf{S}_\mathrm{i}) = \mathbf{S}_\mathrm{i} - \mathrm{MPS}(\mathbf{S}_\mathrm{i}). \tag{4}$$

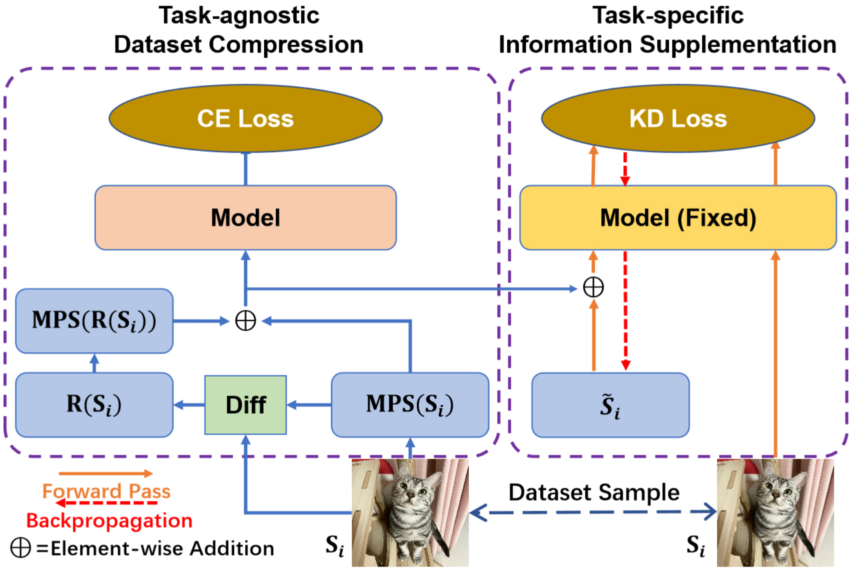

Figure 2: Illustration of the proposed MPSD strategy. $\mathbf{S}_i$ denotes the original image dataset sample. $\mathrm{MPS}(\mathbf{S}_\mathrm{i})$ denotes the MPS decomposed tensor set. $\mathrm{R}(\mathbf{S}_\mathrm{i})$ denotes the difference between $\mathbf{S}_i$ and $\mathrm{MPS}(\mathbf{S}_\mathrm{i})$. $\tilde{\mathbf{S}}_i$ denotes the trainable matrix for distillation. CE loss and KD loss denote the cross-entropy loss function and the knowledge distillation loss function, respectively.

Then, we explicitly let $\mathrm{R}(\mathbf{S}_\mathrm{i})$ approximate the residual information. The original image dataset sample $\mathbf{S}_i$ can be computed as follows:

$$\mathrm{MPS}_{\mathrm{Total}}(\mathbf{S}_i) = \mathrm{MPS}(\mathbf{S}_i) + \mathrm{MPS}(\mathrm{R}(\mathbf{S}_\mathrm{i})). \tag{5}$$

In particular, when $d_k = d_k^{'}$ in $\mathrm{MPS}(\mathbf{S}_\mathrm{i})$, the value of $\mathrm{R}(\mathbf{S}_\mathrm{i})$ is 0 and $\mathrm{MPS}(\mathbf{S}_\mathrm{i})$ is strictly equal to the original matrix $\mathbf{S}_i$ (the upper bound of the error is 0 according to Equation (2)). Algorithm 2 presents a complete procedure for our approach.

Algorithm 2 Task-agnostic Dataset Compression Procedure.

```
Require: Image training dataset with $N$ samples ($\mathbf{S}$).
Ensure: : Compressed training dataset.
for $i = 1 \to N$ do
    Perform MPS decomposition: $\mathrm{MPS}(\mathbf{S}_\mathrm{i}) = \prod_{\mathrm{k}=1}^{\mathrm{n}} \mathcal{T}_{(\mathrm{k})}[\mathrm{d}_{\mathrm{k}-1}, \mathrm{j}_\mathrm{k}, \mathrm{d}_\mathrm{k}]$
    Compress MPS tensors by trucating $\{d_k\}_{k=1}^{n} \longrightarrow \{d_k^{'}\}_{k=1}^{n}$
    Computing residual information: $\mathrm{R}(\mathbf{S}_i) = \mathbf{S}_i - \mathrm{MPS}(\mathbf{S}_i)$
    Perform MPS decomposition: $\mathrm{MPS}(\mathrm{R}(\mathbf{S}_\mathrm{i})) = \prod_{\mathrm{k}=1}^{\mathrm{n}} \mathcal{R}_{(\mathrm{k})}[\mathrm{d}_{\mathrm{k}-1}^{(\mathrm{r})}, \mathrm{j}_\mathrm{k}, \mathrm{d}_\mathrm{k}^{(\mathrm{r})}]$
    Compress MPS tensors by trucating $\{d_k^{(r)}\}_{k=1}^{n} \longrightarrow \{d_k^{'(r)}\}_{k=1}^{n}$
    Computing Compressed sample: $\mathrm{MPS}_{\mathrm{Total}}(\mathbf{S}_i) = \mathrm{MPS}(\mathbf{S}_i) + \mathrm{MPS}(\mathrm{R}(\mathbf{S}_\mathrm{i}))$
end for
return Compressed dataset
```

As a result, we estimate $\mathbf{S}_i$ with multiple MPS. Because each MPS has limited parameters, the sum of their parameters is still considerably smaller than the original $\mathbf{S}_i$, allowing for the dataset compression effect. This idea of using multiple MPS to approximate the original data is inspired by the idea of residuals, and the new MPS state is a description of the discrepancy information between the original data sample $\mathbf{S}_i$ and the truncated representation $\mathrm{MPS}(\mathbf{S}_\mathrm{i})$. We show empirically (Table 1) that our proposed approach can improve the model performance significantly than ICE (Latorre, 2005) for the same compression rate.

4.2 Task-specific Information Supplementation

Task-independent dataset compression is the extraction of short-range correlation information from the dataset. Considering the different network structures and task types, it is necessary to add further information from the learning procedure. Hence we initialize trainable matrix $\tilde{\mathbf{S}}_i$ as implicit bias between $\mathrm{MPS}_{\mathrm{Total}}(\mathbf{S}_{\mathrm{i}})$ and real dataset sample $\mathbf{S}_i$. Then we introduce the knowledge distillation loss function to learn $\tilde{\mathbf{S}}_i$.

Knowledge distillation is used to distill the knowledge from a large training dataset into a small one (Wang et al., 2018). They synthesize data matrix as training data to approximate models trained on the original data. Inspired by data distillation, we initialize trainable matrix $\tilde{\mathbf{S}}_i$ with zeros as implicit bias so that adding $\tilde{\mathbf{S}}_i$ would not hurt the model performance at the first step of training. Similarly, in the context of information supplementation, the synthetic dataset is calculated by $\mathbf{S}_{\mathrm{i}}{}^{*} = \mathrm{MPS}_{\mathrm{Total}}(\tilde{\mathbf{S}}_{\mathrm{i}}) + \tilde{\mathbf{S}}_{\mathrm{i}}$. Then the synthetic dataset $\mathbf{S}_{\mathrm{i}}{}^{*}$ is trained to mimic the behaviors of the real dataset $\mathbf{S}_i$ with the model fixed. Formally, This training process can be modeled as minimizing the following objective function:

$$\mathcal{L}_{\mathrm{KD}} = \sum_{S \in \mathcal{S}} \mathcal{L}(f(\mathrm{MPS}(\mathbf{S}_i) + \mathrm{MPS}(\mathrm{R}(\mathbf{S}_{\mathrm{i}})) + \tilde{\mathbf{S}}_{\mathrm{i}}), \mathrm{f}(\mathbf{S}_{\mathrm{i}})), \tag{6}$$

where $\mathcal{L}(\cdot)$ is a loss function that evaluates the difference between outputs of real and synthetic datasets. Finally, our approach provides a more principle way of information supplementation. By updating implicit bias $\tilde{\mathbf{S}}_i$, the synthetic dataset sample $\mathbf{S}_{\mathrm{i}}{}^{*}$ can better adapt to a specific task or network architecture, and thus achieve better performance.

4.3 The Overall Procedure

Our approach can compress is general. In other words, it can work with existing dataset compression methods to further obtain better compression performance. Here, we choose CIFAR [1] and FashionMNIST [2] as representative image datasets and use our algorithm for these datasets.

The procedure can be simply summarized as follows. First, we perform MPS decomposition separately for the samples in the image dataset. Each sample matrix will be decomposed into a series of local tensors as decrypt in Equation (1). Next, we truncate the connection bond to filtering long-range correlation in the image according to the compression requirements. Then, task-agnostic dataset compression is performed to supplement the difference information between the original image and the truncated MPS image. Finally, a task-specific information supplementation strategy based on knowledge distillation is executed on the task-specific model training. In this way, we expect the dataset to be effectively compressed. In particular, this task-agnostic compressed dataset can be easily generalized to different models as well as to different tasks.

4.4 Discussion

Inspired by Latorre (2005), the image can be represented in the form of matrix product states such that quantum entanglement corresponds to the classical correlation between different coarse-grained textures. The truncation of MPS corresponds to the compression of the original image. In the classification and detection problems, the information that helps the model is mainly short-range correlation information (Krizhevsky et al., 2017). While long-range correlations are noisy for machine learning models and they are not helpful for convergence of machine model training (Gao et al., 2020). With the help of MPS representation of the dataset samples, short-range and long-range correlation information can be effectively decoupling. Our goal is to achieve compression of the dataset by filtering long-range correlation in image. We empirically demonstrate the greater ability of multiple MPS to filter long-range correlation information from images.

In mathematics, MPS-based approximation can be considered as a special low-rank approximation method. We compare it with other low-rank approximation methods, including SVD (Henry & Hofrichter, 1992), CP decomposition (Hitchcock, 1927), and Tucker decomposition (Tucker, 1966). We present the results of these tensor based low-rank approximation methods in Section 5.1 (Table 2).

[1] Available at `https://www.cs.toronto.edu/~kriz/cifar.html`

[2] Available at `https://www.worldlink.com.cn/en/osdir/fashion-mnist.html`

Table 1: The performance comparison to core-set methods and tensor-based methods. This table shows the testing accuracies (%) of different methods on three compression ratios. ResNet18 is used for training and testing. "Whole" indicates an approximate upper-bound performance that is obtained by training the models on the whole dataset.

Datasets	Ratio	ICE	Core-set Selection		Tensor Based		Ours	Whole
			Random	K-Center	SVD	CP		
CIFAR10	40%	89.49	91.67	77.42	87.92	83.94	**92.16**	94.81
	70%	91.67	92.27	79.93	90.47	87.21	**93.32**	
CIFAR100	40%	60.89	62.15	47.63	57.30	51.77	**64.03**	76.48
	70%	64.06	68.76	54.27	62.14	54.24	**70.29**	
FashionMNIST	40%	91.42	91.42	83.36	91.26	89.01	**91.46**	93.83
	70%	92.55	92.78	86.02	91.98	90.11	**93.35**	

In practice, we do not need to strictly follow the original image size. Instead, it is easy to pad additional zero entries to enlarge matrix rows or columns, so that we can obtain different MPS decomposition results. Another note is that the MPS-based approach can work with other compression methods: it can compress matrices condensed by previous methods even more.

5 Experiments

In this section, we first set up the experiments, and then report the results and analysis. Furthermore, the effectiveness of our approach is further demonstrated on other tasks.

Datasets. We first evaluate classification performance with compressed images on four three standard benchmark datasets: CIFAR10, CIFAR100 and FashionMNIST. In particular, the FashionMNIST dataset has 60,000 training and 10,000 testing images of 10 classes, while CIFAR10 and CIFAR100 both have 50,000 training and 10,000 testing images from 10 and 100 object categories, respectively. In all experiments, we use the standard train/test splits of the datasets and finally report the accuracy of the testing dataset.

Baselines. Our baseline methods include:

- ICE (Latorre, 2005): It first transforms images into MPS representation after performing a discrete cosine Fourier transform and truncate the dimensions only once for compression.

- Core-set Selection: It reduces the large dataset into a small equally informative portion of data, including Random and K-Center. In Random, the training samples are randomly selected as the core-set. K-Center (Wolf, 2011) picks multiple center points such that the largest distance between a data point and its nearest center is minimized.

- Tensor Based Methods: They compress the dataset by applying low-rank approximation to each image, including Tucker decomposition and CP decomposition.

Implementations. Following Latorre (2005), we first represent the image with MPS format (*i.e.*, a product of n local tensors) and then apply task-agnostic dataset compression as well as task-specific information supplementation to reduce total dataset size. To ensure a fair comparison, we adopt the same architecture, ResNet18 (He et al., 2016), , for different dataset compression methods. For simplicity and generality, we use default model hyper-parameters of ResNet18 and a consistent augmentation strategy for all datasets. For example, the learning rate, the minibatch size, and training epochs are set to 0.1, 128 and 180, respectively. Finally, we report results on testing datasets with the best model on evaluation datasets.

5.1 Experimental Results

Comparison to Image Compression. As shown in Table 1, our approach is very competitive in the three image classification benchmark, and it outperforms the image compression method in all

Table 2: Evaluations on different network architectures. This table shows the testing accuracy (%) of different methods at a compression ratio of 70%.

Datasets	ResNet18		VGG		MobileNetV2	
	ICE	Ours	ICE	Ours	ICE	Ours
CIFAR10	89.49	92.04	88.07	92.21	85.91	89.56
CIFAR100	64.06	70.29	60.80	67.71	61.05	66.39
FashionMNIST	92.55	93.35	92.37	93.21	92.73	93.93

tasks. Looking at CIFAR100, compared with ICE, our approach improves 6.23% in terms of test performance at the same compression ratio. By zooming in on a specific dataset, the improvements over CIFAR100 are larger than the other tasks. Note that compared to CIFAR10 and FashionMNIST, CIFAR100 is more challenging, as recognizing 10 times more categories with $\frac{1}{10}$ fewer images per class in CIFAR100. The MPS dataset seems to work better with few shot tasks, which enhances the data efficiency of the training dataset.

Comparison to Core-set Methods. To demonstrate the strength of image compression with MPS over the core-set selection, we do experiments on CIFAR10 and use Random and K-Center for comparison. Table 1 summarizes the results. Overall, our approach achieves competitive results over core-set selection methods, especially for the K-Center method. This is achieved due to the completeness that the dataset compression method has and its intrinsic characteristic of effectively preserving short-range correlation for each image. To compare the quality of truncated MPS representation for CIFAR10 and understand them intuitively, we visualize images from five categories with different dimensions $d_k^{'}$ in Figure 3. We observe that it is impossible to see the difference before and after compression if ρ is larger than 36%. Compared to losing some images by the core-set selection, filtering long-range correlation information by our approach can minimize the damage to the dataset.

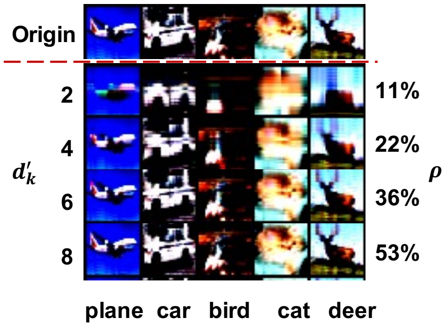

Figure 3: Illustration of low-rank approximation for MPS to CIFAR10 images. $d_k^{'}$ and ρ denotes truncated dimension of $\mathrm{MPS}(\mathbf{S})$ and compression ratio, respectively.

Comparison to Tensor Based Methods. As discussed in Section 4.4, we use other tensor-based methods (*i.e.,* SVD (Henry & Hofrichter, 1992), CP (Hitchcock, 1927)) for comparison to demonstrate the effectiveness of preserving short-range correlations with MPS. From the Tabel 1, we observe that SVD and CP decomposition failed to preserve useful information for model performance especially when the compression ratio is less than 40%. While our approach can still have competitive accuracy over CIFAR10 when very limited information is preserved.

Comparison to Different Models In general, our approach can be applied to any kind of network architecture. We have evaluated its performance with ResNet18. Now, We continue to test our approach using another two standard deep network architectures: VGG-16 (Simonyan & Zisserman, 2015) and MobileNetV2 (Sandler et al., 2018). These models are famous pre-trained models that showed state-of-the-art accuracy for several challenging recognition tasks on ImageNet and competitions. Table 2 presents the comparison of the testing accuracy with three network architectures. As we can see, the MPS dataset can cooperate with different kinds of network architectures.

5.2 Evaluation on More tasks

As introduced in Section 4, our approach contains task-agnostic dataset compression and task-specific information supplementation. Due to task-agnostic compression, MPS representation can be applied in other computer vision tasks (*i.e., Pedestrian Detection*, *Visual Question Answering* and *Large-scale Image Classification*).

Pedestrian Detection First, we apply our approach to an pedestrian detection scenario where the goal is to accurately locate pedestrians in an image. We build our model on Mask R-CNN (He et al., 2017) method and fine-tune a pre-trained Mask R-CNN model in the Penn-Fudan Database (Wang et al., 2007) for Pedestrian Detection and Segmentation task. The dataset contains 170 images with 345 instances of pedestrians and we decompose the original images with MPS and use a trainable matrix to supplement information. The desired outcome is to obtain a high mean of average precision over all the classes. Finally, we report a COCO-style mAP score after 10 epochs of training, and the result is shown in Table 3. The result indicates that MPS dataset can achieve competitive model performance while reducing 25% parameters of the dataset.

Visual Question Answering The Visual Question Answering task typically uses paired images and text to bridge vision and language respectively. Current approaches to this heavily rely on image feature extraction processes. Here we explore the use of our approach on VQAv2. The VQAv2 task asks for answers given pairs of an image and a question in natural language. Test-dev score is calculated by comparing the inferred answer to the 10 ground-truth answers. Our goal is to verify that short-range correlation in images can be used to efficiently model the cross-modal interaction between image-text pairs. To this end, we replace the image with MPS representation in the image-text pairs. Following Kim et al. (2021), we use a pre-trained ViLT model and fine-tune the model on the MPS dataset. Finally, from Tabel 3 we observe that our approach achieves comparable testing performance, and meanwhile significantly decreases the size of the dataset.

Large-scale Image Classification Pre-trained deep learning models (ResNet, VGG) learned on large-scale datasets have shown their effectiveness over conventional methods. Instead of training a model from scratch, one can fine-tune a pre-trained model to solve some specific task. To demonstrate the effectiveness of short-range correlation on transfer learning, we apply our approach to the ImageNet dataset. To this end, we observe that the total dataset size is significantly reduced due to the compression on each image in the dataset. Furthermore, we evaluate the performance of the pre-training ResNet18 model on both the original ImageNet and the MPS compressed dataset. We observe that the MPS dataset achieves comparable accuracy to the original ImageNet dataset. This result shows that the MPS dataset with short-range correlation can support large-scale pre-training.

Table 3: The performance comparison with the ICE method, both our proposed approach and the ICE method have a compression ratio of 75%.

Experiments	**Object Detection**	**Multimodal Task**	**Large Scale Pre-training**
	Pedestrian Detection (mAP)	VQAv2 (test-dev score)	ImageNet (acc)
Origin	79.90	70.33	64.21
ICE	67.32	57.27	47.13
Ours	**73.45**	**63.78**	**52.10**

6 Conclusion

We propose a dataset compression approach based on MPS and distillation. With MPS decomposition, it is able to efficiently reorganize and decouple short-range and long-range correlation information in local tensors. The MPS can be used to correspond the classical correlations to quantum entanglement, and the short-range and long-range correlations to small and large magnitude of entanglement entropy, respectively. We empirically found that the short-range correlation in images is important for training. Inspired by this, we design a novel dataset compression approach that achieves effective compression of dataset by filtering long-range correlation features from images in task-agnostic scenario, while using distillation to complement task-relevant information. Extensive experiments have demonstrated the effectiveness of our approach, especially in that the compressed dataset using the MPS decomposition can be directly applied to a variety of different neural network tasks. To the best of our knowledge, this is the first application of MPS for dataset compression. In future work, we will consider exploring more decomposition structures for MPS.

References

Pankaj K. Agarwal, Sariel Har-Peled, and Kasturi R. Varadarajan. Approximating extent measures of points. *J. ACM*, 51(4):606–635, 2004.

Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 11816–11825, 2019.

Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 2654–2662, 2014.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 535–541, 2006.

Pasquale Calabrese and John Cardy. Entanglement entropy and quantum field theory. *Journal of statistical mechanics: theory and experiment*, 2004(06):P06002, 2004.

Francisco M. Castro, Manuel J. Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XII*, volume 11216 of *Lecture Notes in Computer Science*, pp. 241–257. Springer, 2018.

Ke Chen. On coresets for k-median and k-means clustering in metric and euclidean spaces and their applications. *SIAM J. Comput.*, 39(3):923–947, 2009.

Andrzej Cichocki, Rafal Zdunek, Anh Huy Phan, and Shun-ichi Amari. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019.

Mark Fannes, Bruno Nachtergaele, and Reinhard F Werner. Finitely correlated states on quantum spin chains. *Communications in mathematical physics*, 144(3):443–490, 1992.

Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca, and projective clustering. *SIAM J. Comput.*, 49(3):601–657, 2020.

Ze-Feng Gao, Song Cheng, Rong-Qiang He, Z. Y. Xie, Hui-Hai Zhao, Zhong-Yi Lu, and Tao Xiang. Compressing deep neural networks by matrix product operators. *Phys. Rev. Research*, 2:023300, Jun 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 2980–2988. IEEE Computer Society, 2017.

ER Henry and J Hofrichter. [8] singular value decomposition: Application to analysis of experimental data. *Methods in enzymology*, 210:129–192, 1992.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.

Frank L Hitchcock. The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics*, 6(1-4):164–189, 1927.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5583–5594. PMLR, 2021.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, 2017.

José Ignacio Latorre. Image compression and entanglement. *CoRR*, abs/quant-ph/0510031, 2005.

Alexander Novikov, Dmitry Podoprikhin, Anton Osokin, and Dmitry P. Vetrov. Tensorizing neural networks. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 442–450, 2015.

D Perez-Garcia, Frank Verstraete, MM Wolf, and JI Cirac. Matrix product state representations. *QUANTUM INFORMATION & COMPUTATION*, 7(5-6):401–430, 2007.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 5533–5542. IEEE Computer Society, 2017.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

MohamedAR SainathTN et al. Deep convolutionalneuralnetworksforlvcsr. *IEEE InternationalConferenceon Acoustics, Speechand SignalProcessing*, 8614:8618, 2013.

Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 4510–4520. Computer Vision Foundation / IEEE Computer Society, 2018.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Mark Srednicki. Entropy and area. *Phys. Rev. Lett.*, 71:666–669, Aug 1993.

Xingwei Sun, Ze-Feng Gao, Zhong-Yi Lu, Junfeng Li, and Yonghong Yan. A model compression method with matrix product operators for speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2837–2847, 2020.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

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

Guifré Vidal. Efficient simulation of one-dimensional quantum many-body systems. *Physical review letters*, 93(4):040502, 2004.

Liming Wang, Jianbo Shi, Gang Song, and I-Fan Shen. Object detection combining recognition and segmentation. In Yasushi Yagi, Sing Bing Kang, In-So Kweon, and Hongbin Zha (eds.), *Computer Vision - ACCV 2007, 8th Asian Conference on Computer Vision, Tokyo, Japan, November 18-22, 2007, Proceedings, Part I*, volume 4843 of *Lecture Notes in Computer Science*, pp. 189–199. Springer, 2007.

Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A. Efros. Dataset distillation. *CoRR*, abs/1811.10959, 2018.

Gert W. Wolf. Facility location: concepts, models, algorithms and case studies. series: Contributions to management science. *Int. J. Geogr. Inf. Sci.*, 25(2):331–333, 2011.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Appendix

We offer some proof and details of experiments to help audience better understand our approach. The appendix includes 2 pages and is organized into sections:

- Tensor and Matrix Product States
- Theorem
- Experiment

A Tensor and Matrix Product States

As introduced in Cichocki et al. (2009), the concept of tensor is specified as:

Definition1
(Tensor). Let $D_1, D_2..., D_N \in N$ denote index upper bounds. A tensor $\mathcal{T} \in \mathbb{R}^{D_1,...,D_n}$ of order N is an N-way array where elements $\mathcal{T}_{d_1,d_2,...,d_n}$ are indexed by $d_n \in \{1, 2, ..., D_n\}$ for $1 \leq n \leq N$

Definition2
(Matrix product state). We can reshape a matrix to high order tensor, denote as:

$$\mathbf{M}_{x\times y} = \mathbf{M}_{i_1 i_2 ... i_n}. \tag{7}$$

Here, the one-dimensional coordinate x and y means the input and output signal, respectively. So that $\mathbf{x} \times \mathbf{y}$ with dimension $N_x N_y$ is reshaped into a coordinate in a n-dimensional space, labelled by $(i_1 i_2 \cdots i_n)$. Hence, there is a one-to-one mapping between original matrix and $(i_1 i_2 \cdots i_n)$. If I_k are the dimensions of i_k, then:

$$\prod_{k=1}^{n} I_k = N_x N_y. \tag{8}$$

The MPS representation of $\mathbf{M}$ is obtained by factorizing it into a product of n local tensors.

$$M_{i_1 \cdots i_n, j_1 \cdots j_n} = \mathcal{T}^{(1)}[i_1, d_1, j_1] \cdots \mathcal{T}^{(n)}[i_n, d_n, j_n], \tag{9}$$

where $\mathcal{T}^{(k)}[i_k, d_k, j_k]$ is a 3-order tensor. The D_k is the virtual basis dimension on the bond linking $\mathcal{T}^{(k)}$ and $\mathcal{T}^{(k+1)}$ with $D_0 = D_n = 1$.

$D_{k-1} \times D_k$ matrix with D_k the virtual basis dimension on the bond linking $\mathcal{T}^{(k)}$ and $\mathcal{T}^{(k+1)}$ with $D_0 = D_n = 1$.

B Theorem

Theorem 1. Suppose that the tensor $\mathbf{W}^{(k)}$ of matrix W that is satisfy

$$\begin{aligned} &\mathbf{W} = \mathbf{W}^{(k)} + \mathbf{E}^{(k)}, D(\mathbf{W}^{(k)}) = d_k, \\ &where \quad ||\mathbf{E}^{(k)}||_F^2 = \epsilon_k^2, k = 1, ..., d-1. \end{aligned} \tag{10}$$

Then MPS $(\mathbf{W})$ with the k-th bond dimension d_k upper bound of truncation error satisfy:

$$||\mathbf{W} - \mathrm{MPS}\ (\mathbf{W})||_{\mathrm{F}} \leq \sqrt{\sum_{\mathrm{k=1}}^{\mathrm{d-1}} \epsilon_{\mathrm{k}}^2} \tag{11}$$

Proof. The proof is by induction. For $n = 2$ the statement follows from the properites of the SVD. Consider an arbitrary $n > 2$. Then the first unfolding $\mathbf{W}^{(1)}$ is decomposed as

$$\mathbf{W}^{(1)} = \mathbf{U}_1 \lambda_1 \mathbf{V}_1 + \mathbf{E}^{(1)} = \mathbf{U}_1 \mathbf{B}^{(1)} + \mathbf{E}^{(1)} \tag{12}$$

where $\mathbf{U}_1$ is of size $r_1 \times i_1 \times j_1$ and $||\mathbf{E}^{(1)}||_F^2 = \epsilon_1^2$. The matrix $\mathbf{B}_1$ is naturally associated with a $(n-1)$-dimensional tensor $\mathcal{B}^{(1)}$ with elements $\mathcal{B}^{(1)}(\alpha, i_2, j_2, ..., i_n, j_n)$, which will be decomposed further. This means that $\mathbf{B}_1$ will be approximated by some other matrix $\hat{\mathbf{B}_1}$. From the properties of the SVD it follows that $\mathbf{U}_1^T\mathbf{E}^{(1)} = 0$, and thus

$$\begin{aligned}
&||\mathbf{W} - \mathcal{B}^{(1)}||_F^2 \\
&= ||\mathbf{W}_1 - \mathbf{U}_1\hat{\mathbf{B}_1}||_F^2 \\
&= ||\mathbf{W}_1 - \mathbf{U}_1(\hat{\mathbf{B}_1} + \mathbf{B}_1 - \mathbf{B}_1)||_F^2 \\
&= ||\mathbf{W}_1 - \mathbf{U}_1\mathbf{B}_1||_F^2 + ||\mathbf{U}_1(\hat{\mathbf{B}_1} - \mathbf{B}_1)||_F^2
\end{aligned} \tag{13}$$

and since $\mathbf{U}_1$ has orthonormal columns,

$$||\mathbf{W} - \mathcal{B}^{(1)}||_F^2 \leq \epsilon_1^2 + ||\mathbf{B}_1 - \hat{\mathbf{B}_1}||_F^2. \tag{14}$$

and thus it is not difficult to see from the orthonormality of columns of $\mathbf{U}_1$ that the distance of the k-th unfolding $(k = 2, ..., d_k - 1)$ of the $(d-1)$-dimensional tensor $\mathcal{B}^{(1)}$ to the d_k-th rank matrix cannot be larger then ϵ_k. Proceeding by induction, we have

$$||\mathbf{B}_1 - \hat{\mathbf{B}_1}||_F^2 \leq \sum_{k=2}^{d-1} \epsilon_k^2, \tag{15}$$

combine with Eq. equation 14, this complets the proof.

C Experiment

C.1 Additional Details of MPS Dataset

In this paper, the MPS is proposed for compressing CIFAR10, CIFAR100 and FashionMNIST, respectively. In order to show that the process of incorporating several MPS structures into different dataset. We introduce MPS decomposition in different image shape:

Layers	Compression Ratio	MPS shape $[d_{k-1}, i_k, j_k, d_k]$	Compression Ratio	MPS shape $[d_{k-1}, i_k, j_k, d_k]$
CIFAR10,CIFAR100	71%	$\mathcal{T}_1 : [1, 4, 1, 4]$ $\mathcal{T}_2 : [4, 4, 1, 6]$ $\mathcal{T}_3 : [6, 4, 1, 6]$ $\mathcal{T}_4 : [6, 4, 1, 4]$ $\mathcal{T}_5 : [4, 4, 1, 1]$	43%	$\mathcal{T}_1 : [1, 4, 1, 4]$ $\mathcal{T}_2 : [4, 4, 1, 2]$ $\mathcal{T}_3 : [2, 4, 1, 2]$ $\mathcal{T}_4 : [2, 4, 1, 4]$ $\mathcal{T}_5 : [4, 4, 1, 1]$
FashionMNIST	66%	$\mathcal{T}_1 : [1, 2, 1, 2]$ $\mathcal{T}_2 : [2, 4, 1, 6]$ $\mathcal{T}_3 : [6, 7, 1, 5]$ $\mathcal{T}_4 : [5, 7, 1, 2]$ $\mathcal{T}_5 : [2, 2, 1, 1]$	33%	$\mathcal{T}_1 : [1, 2, 1, 2]$ $\mathcal{T}_2 : [2, 4, 1, 6]$ $\mathcal{T}_3 : [6, 7, 1, 2]$ $\mathcal{T}_4 : [2, 7, 1, 2]$ $\mathcal{T}_5 : [2, 2, 1, 1]$
ImageNet	63%	$\mathcal{T}_1 : [1, 7, 1, 7]$ $\mathcal{T}_2 : [7, 8, 1, 28]$ $\mathcal{T}_3 : [28, 16, 1, 28]$ $\mathcal{T}_4 : [28, 8, 1, 7]$ $\mathcal{T}_5 : [7, 7, 1, 1]$	38%	$\mathcal{T}_1 : [1, 7, 1, 7]$ $\mathcal{T}_2 : [7, 8, 1, 21]$ $\mathcal{T}_3 : [21, 16, 1, 21]$ $\mathcal{T}_4 : [21, 8, 1, 7]$ $\mathcal{T}_5 : [7, 7, 1, 1]$

Table 4: MPS structures of CIFAR10, CIFAR100, FashionMNIST and ImageNet datasets.

We design different MPS structures based on the image size of the dataset. Accordingly, the compression ratio is calculated by corresponding truncated dimension.