On Approximating the Non-negative Rank: Applications to Unsupervised Image Reduction

Mohanad Abukmeil, Stefano Ferrari, Angelo Genovese, Vincenzo Piuri, and Fabio Scotti
Department of Computer Science
Università degli Studi di Milano
Milan, Italy
{mohanad.abukmeil, stefano.ferrari, angelo.genovese, vincenzo.piuri, fabio.scotti}@unimi.it

***Abstract*—Unsupervised Learning (UL) methods are a class of machine learning which aims to disentangle the representations and reduce the dimensionality among the data without any predefined labels. Among all UL methods, the Non-negative Matrix Factorization (NMF) factorizes the data into two subspaces of non-negative components. Moreover, the NMF enforces the non-negativity, sparsity, and part-based analysis, thus the representations can be interpreted and explained for the Explainable Artificial Intelligence (XAI) applications. However, one of the main issues when using the NMF is to impose the factorization rank *r* to identify the dimensionality of the subspaces, where the rank is usually unknown in advance and known as the non-negative rank that is used as a prior to carrying out the factorization. Accordingly, we propose a novel method for the non-negative rank r approximation to help solving this problem, and we generalize our method among different image scales. Where, the results on different image data sets confirm the validity of our approach.**

***Index Terms*—Unsupervised Learning, Explainable Artificial Intelligence (XAI), Non-negative Matrix Factorization, Non-negative Rank, Image Reduction.**

I. INTRODUCTION

Explainable Artificial Intelligence (XAI) is considered an emerging field in Machine Learning (ML), with the aim of interpreting how the decisions of ML applications are made [1], [2]. The early stages of any XAI model include data reduction and disentangling the interpretable representations, *i.e.*, the explainable features that reflect parts of the data [3].

To disentangle the representations and reduce data dimensionality, Unsupervised Learning (UL) methods have been especially studied, due to their advantage of working on unlabeled data. In particular, the UL methods include Principal Component Analysis (PCA), Independent Component Analysis (ICA), Autoencoder (AE), Non-negative Matrix Factorization (NMF), Tensor Decomposition (TD), and others [4]–[6]. Among all UL methods, the NMF decomposes the data into low dimensional subspaces, where the first one contains the bases of the latent features and is named as the latent space W, the other space hides the coefficients that reconstruct the data and is called the mixing space H [7]. The capability of the NMF lies in the fact that it factorizes the data into non-negative components, which usually in low dimensional form and represent portions of the original data itself. Moreover, the other advantage of the NMF is that it is an inherently sparse method for feature representations, thus it offers sparse components (subspaces) that are isolated and represent the original data objects [8].

The NMF can be integrated into the XAI models to improve their explainability and interpretability, as it is easy to relate the hidden (or latent) representations to the original data itself. Also, it reduces the dimensionality and the computational time required to disentangle the interpretable representations among the data. However, one of the main issues when using the NMF is to identify the factorization rank r among the data, which is usually unknown in advance and can be imposed as a prior to carrying out the factorization [9].

To help solving this problem, we propose a novel rank approximation method for the NMF, where our contribution is twofold: *(I)* introducing an effective rank truncation method based on a mathematical operators and a physical conceive, thus the dimensionality among the data and the processing time are reduced according to the approximated rank; *(II)* generalizing the rank truncation from small-size data samples to a large-size ones, thus achieving adaptability at different scales. The rest of this paper is organized as follows: Section 2 highlights the β-NMF. Section 3 outlines the NMF rank analysis. The experimental results are reported in Section 4. Section 5 summarizes the conclusion and future works.

II. THE β-NMF FACTORIZATION

For a given image $X \in \mathbb{R}^{m \times n}$, the NMF decomposes the data matrix as $X \approx WH + R_s$, where $W \in \mathbb{R}^{m \times r}$ contains the bases of the latent subspace, $H \in \mathbb{R}^{r \times n}$ represents the coefficients of the mixing subspace, R_s is the residual, and r is the input rank. The NMF expresses the data as a product of two subspaces, where the factorization is carried out by minimizing the objective function that measures the mismatch between the original image data and the reconstructed subspaces [7]. Specifically, the β-divergences is a class of the objective functions comprising the Itakura-Saito (IS) when $\beta = 0$, Kulback Leibler (KL) when $\beta = 1$, and Frobenius norm when $\beta = 2$. These objective functions are used to quantify the distance between the original image and the reconstructed one obtained from the factorized subspaces, *i.e.*, W and H [10], [11]. The β-divergence between two matrix elements is given

as:

$$d_\beta(x,\tilde{x}) = \begin{cases} \frac{x}{\tilde{x}} - \log\frac{x}{\tilde{x}} - 1, & \beta = 0 \\ x\log\frac{x}{\tilde{x}} - x + \tilde{x}, & \beta = 1 \\ \frac{1}{\beta}(\beta-1)(\tilde{x}^\beta + (\beta-1)\tilde{x}^\beta - \beta x\tilde{x}^\beta), & \text{otherwise} \end{cases} \quad (1)$$

where *d* is the divergence, *x* represents the original image data pixel (or point), $\tilde{x}$ is the reconstructed pixel after applying the factorization or learning. When extending the notation from pixel or data point to matrix (whole image), the β-divergence generalization is given as:

$$d_\beta(X,\tilde{X}) = \sum_{(i,j)} d_\beta(X_{(i,j)}, (W_r H_r)_{(i,j)}) \quad (2)$$

where d_β is the divergence, X is the original image data, $\tilde{X} = W_r H_r$ are the bases of the latent subspace and coefficients of the mixing subspace, respectively, and resulting by the factorization using rank r. Furthermore, the matrix update for the bases W and coefficients H (*i.e.*, changing the latent factors for W and H to reach the minimum divergence) can be done according to the following rules [12]:

$$W \leftarrow W \odot \frac{([WH]^{\odot\beta-2} \odot X)H^{\mathrm{T}}}{[WH]^{\odot\beta-1}H^{\mathrm{T}}} \quad (3)$$

$$H \leftarrow H \odot \frac{W^{\mathrm{T}}([WH]^{\odot\beta-2} \odot X)}{W^{\mathrm{T}}[WH]^{\odot\beta-1}} \quad (4)$$

where T is the matrix transpose and $\odot$ denotes the element-wise multiplication.

The most widely used approaches for the NMF initializations are SVD-NMF, non-negative double SVD, and non-negative SVD with low-rank correction [13]. These methods share a similar procedure to carry out the NMF initialization and identifying the dimensions of both W and H spaces; by updating the matrices W and H based on the rank increment at each iteration until a given level of the performance is reached. Other approaches, based on the rank adaptation from the lowest to the highest or full rank, are time-consuming since they depend on a trial and error procedure [14].

To avoid the cost given by the iterative search, a commonly used practical approach (the rule of thumb) keeps the singular values that contribute to $90:99\%$ of the total energy sum and impose their number as the rank, however, such an approach suffers from the instability due to fixing the bounds of the singular values [15]. The recent method proposed in [16] based on the Minimum Description Length (MDL). The MDL method depends on finding a possible way to encode the data with high precision and low decoding error, where it does not approximate the rank directly from the data itself; it selects the suitable rank that reflects the minimum MDL among a list of all imposed ranks, thus it can be seen as a kind of trial and error method. Moreover, the MDL method assumes that the data samples are already factorized with all available ranks and the subspaces W and H which correspond to each rank are kept, then it converts the subspace to distributions and utilizes the Shannon information content [17] among the distributions to measure the minimum MDL that reflect the best rank.

To obtain the factorization rank automatically, we will propose an innovative rank approximation based on the combinations between the matrix trace, nuclear-, and the Frobenius-norm. Such a combination able to preserve suitable bounds of the singular values that reflect the optimal rank, also it obtains the rank directly from the data itself to achieve the stability and reduce the computational cost.

III. PROPOSED RANK ANALYSIS

Advocated by the fact that the NMF requires to impose the factorization rank *r* before carrying out the factorization, *i.e.*, the size of the data columns and row spaces that identify the dimensions of the NMF subspaces, also the rank must be accurate and able to preserve an acceptable level of the reconstruction accuracy [13]. Practically, running the empirical rank approximations for the NMF factorization among all available ranks is considered a computational burden, especially when the data comes in a multi-way form as in the RGB images and when its size is relatively big [8]. For instance, to identify the optimal rank for an image with dimensions of 256×256 empirically, we need to run 256 iterations from the lower rank $(r = 1)$ to the full rank $(r = 256)$. Thus, to reach the appropriate rank without alternating all possible ranks, we propose an automatic rank truncation procedure, which can be useful for different data sets and based on the linear transformations among the data.

Both the trace and nuclear norm have been utilized to produce a very low-rank solution theoretically [18], [19], however, we extend the theory to the practice by proposing a suitable rank approximation for ML data sets. The trace of the data matrix (the data matrix is an image and denoted as X in all parts of this work) is considered a useful linear transformation, and it gives the derivative of the determinant $\det|X|$ that offers the volume equipped by the column space to approximate the data rank [20]. Mathematically, the trace is the sum of all diagonal elements of a square matrix (or sum of the eigenvalues), whereas the physical meaning of the trace is the constructions of the Hamiltonians (the total energy) of the quantum system that associated with a finite set of the energy eigenvalues [21]. The trace of the $n \times n$ square matrix X is given as:

$$\mathrm{Tr}(X) = \sum_{i=1}^{n} x_{(i,i)}, \qquad i = 1,,,n \quad (5)$$

In the same orientation, the nuclear norm $||\cdot||_*$ is considered a substantial tool in the field of multivariate statistics and dimensionality reduction, whereas it used recently for deep learning optimization as a convex replacement of the dimensional rank [19]. The nuclear norm reflects the importance of the singular values that constitute the rank variation among the data. The nuclear norm can be obtained by summing all singular values, which can be retrieved using the Singular Value Decomposition (SVD) [22]. Where the singular values' matrix is constituted in a diagonal form and reflects the importance of the eigenvalues and data points when reconstructing the data. Factually, summing all singular values is equivalent to adding

up the absolute values of the diagonal elements (*i.e.* the L_1 norm) of the diagonal matrix, where the rank minimization problem is tuned to find a sparse vector in the affine subspace [19]. The nuclear norm of a given data matrix X is obtained from the SVD as:

$$X \approx \sum_{i=1}^{r} U_i \sigma_i V_i^{\mathrm{T}}, \quad \text{where} \quad ||X||_* = \sum \sigma_i \tag{6}$$

where U_i and V_i are orthonormal matrices called the left and right singular vectors, respectively, and they represent the eigenvectors of $XX^{\intercal}$ and $X^{\intercal}X$, respectively. Additionally, σ_i is a diagonal matrix that contains the singular values $\{\sigma_i \,|\, i = 1, \ldots, n\}$ of the matrix X and are sorted in decreasing order.

In our proposed approach, we combine the Frobenius norm [23] to penalize the rank truncation threshold, due to its capability to the transformations which are unitary invariant. Thus, the bound of the singular values can be increased to an appropriate level reaching the optimal rank. The Frobenius norm for a given matrix $X \in \mathbb{R}^{m \times n}$ can be obtained by the square root of the sum of the squares of the matrix elements as follows:

$$||X||_{\mathrm{F}} = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} x_{(i,j)}^2} \tag{7}$$

The nuclear norm and the trace of the data are considered algebraic tools, which help to identify a truncation limit to remove the unwanted portions of data and only keep the ones that reflect the rank r. The rank truncation limit is a threshold ϵ at which performance can be saturated and the data can be reduced [20], [24]. Dividing the trace of eigenvalues by the nuclear norm (as in Eqn. 8) is equivalent to spreading out the total energy among different separate states with different capacities, where the highest singular values absorb the highest energies and the lowest values only get very low energies [21].

Because the trace of the square matrix is equal to the sum of its eigenvalues[1], there is an ability to compute the trace directly from the data matrix without the need to diagonalize the data (in Eqn. 8) to obtain the trace of its eigenvalues matrix. However, if the data matrix in a non-square form or if all values are zeros in its diagonal, it must diagonalize the matrix first (*i.e.,* to covert it as sets of eigenvalues and eigenvectors), then summing of its eigenvalues as a trace.

The proposed threshold is able to identify suitable bounds of singular values that reflect the rank; by truncating the singular values that absorb the minimum energies, and only keeping the highest ones which conclude the essential and appropriate features among the data. Our methodology is divided into four steps as follows:

- For a given image $X \in \mathbb{R}^{n \times n}$, calculate the $\mathrm{Tr}(X)$ and $||X||_{\mathrm{F}}$.
- Calculate the SVD and obtain $||X||_*$.
- Identify the truncation threshold as:

$$\epsilon = \frac{\sqrt{\mathrm{Tr}(X)}}{||X||_* + ||X||_{\mathrm{F}}} \tag{8}$$

- The rank is obtained by counting the singular values (taken from SVD) up to ϵ.

Accordingly, the rank can be expressed as the total number of the singular values that lie up to the truncation threshold ϵ.

To measure the performance of our proposed method, we will employ both MNIST digits and MNIST fashion data sets [25], [26], which share similar small image dimensions. Thus, there is a possibility to build a common threshold to obtain the non-negative rank as it appears in the Eqn. 8. Moreover, we found that it is important to inject the Frobenius norm to the threshold denominator as a penalty factor, to expand the minimum bound of singular values into an appropriate limit that reflect the optimal rank.

However, for the data sets which are relatively larger than the MNIST on terms of image size, the trace and the nuclear norm can be fixed without injecting the Frobenius norm; since the MNIST data sets contain a lot of zero elements in the images which affect the bounds of the singular values. Following the same footprint in approximating the rank for the MNIST data sets, we can build similar rank truncation based on the trace and the unclear norm for the color and large-sized images. The rank truncation threshold for such type of images that hides a lot of information due to their size and dimensionality can be designed as:

$$\epsilon = \sqrt{\frac{\mathrm{Tr}(X)}{||X||_*}} \tag{9}$$

IV. EXPERIMENTAL RESULTS

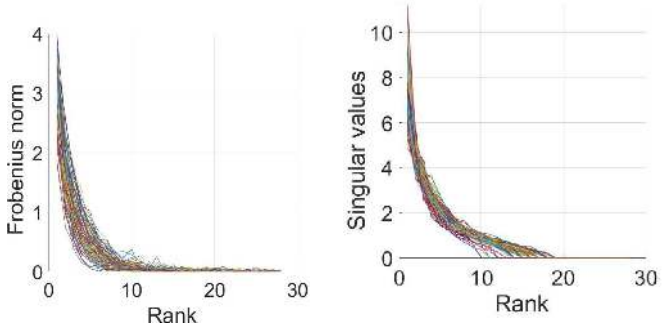
A. Approximating the rank for small-size images

The MNIST digits data set comprises 70000 handwritten images divided in 10 classes, representing digits from $0-9$. The kin new data set to the MNIST digits is called the MNIST fashion with the same number of samples. The size of each image in the MNIST data sets is similar, and each image size is 28×28 which forms the data columns and the rows spaces. The MNIST data sets are ideal for analyzing the small-size images, due to the limited size of the images. Where our used computational platform to approximate the NMF rank includes Intel CPU Core $i7{-}4800MQ$ processor with 8G DDR3 RAM, and all experiments have carried out using MATLAB R2019b on windows 10 operating system.

To extract the non-negative features for the MNIST data sets, the factorization rank r among the images must be imposed. Traditionally, the rank can be approximated by alternating the rank from $r = 1$ to the full rank-size where $r = 28$. We used the β-NMF [11] with $\beta = 1$ to show how the traditional rank affects the image reconstruction. Moreover, we used the Frobenius norm as a function of image rank in the following Fig. 1(a) to depict how the original images and the reconstructed ones are similar to each other. Furthermore, to show the relationship between the rank approximation and the

[1] *http : //people.math.harvard.edu/ ∼ knill/teaching/ math19b_2011/handouts/math19b_2011.pdf*

distribution of singular values of the same data used in Fig. 1(a), we plot the singular values and the rank (from $r = 1 : 28$) for each image as it can be illustrated in Fig. 1(b). For the sake of space, we plot only the Frobenius norm and singular values distributions for the first 100 images of the third class from the MNIST digits, where the idea is generalized among all samples and classes.



(a) (b)

Fig. 1. (a) The Frobenius norm of 100 images of the third class as a function of rank r, (b) the corresponding singular values of the same images in (a). It is possible to observe that the Frobenius norm decays at $r = 15$, where the optimal bound of the singular values lies at the same rank.

As can be noticed from the Fig. 1(a) and Fig. 1(b) when the rank is very low at $r = 1 : 5$, the Frobenius norm is very high, conversely, the norm is saturated in the pool of ranks greater than 15. The rule of thumb [15] to extract the rank from the data can be carried out by retaining $90 - 99\%$ of the singular values that contribute to the sum of total energies and utilizing their number as the NMF rank. However, according to Fig. 2 using the rule of thumb reflects the rank $r = 10$, also at that rank the averaged Frobenius norm between the original images and the reconstructed ones for the same samples used in Fig. 1(a) and Fig.1 (b) is equal to 0.0209. Whereas, our method expands the bounds of the singular values to enhance the performance and gives the rank $r = 15$, also the reconstruction loss obtained by the Frobenius norm enhanced and reached to 0.0049 when comparing the useful rule of thumb [15].

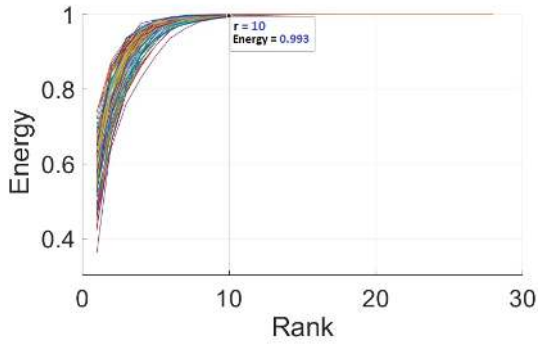


Fig. 2. The distribution of the accumulated energies of the singular values. It is apparent from the figure that the largest $r = 10$ singular values contribute for the 99% of their total energy.

Another used approach in the practice, by keeping 90% of the singular values that contribute to the nuclear norm, where it gives the average rank for the images used in Fig. 1 equal to $r = 9$. Also, at that rank the Frobenius norm still not saturated, *i.e.*, the reconstruction loss requires to increase the rank to be enhanced. From this angle, it is important to build a suitable rank threshold to be utilized for automatic non-negative rank approximation, thus the rank can be imposed easily with minimum reconstruction loss as what we proposed in Eqn. 8.

Table I shows the average rank for each class extracted from the MNIST data sets using our threshold in Eqn. 8, using only the average rank of 1000 random images from each class to be generalized among the whole images within each class. Moreover, the table presents the performance analysis of all data sets samples (not just 1000 samples) when using the generalized rank, utilizing the Frobenius norm and the SSIM index to compare the original images and the reconstructed ones after the factorization. For more details about the SSIM index we refer to [27].

TABLE I
THE APPROXIMATED RANK AND THE PERFORMANCE ANALYSIS FOR THE MNIST DATA SETS.

MNIST Digit					MNIST Fashion				
ClassID	**#Sample**	**Rank**	**Fnorm**	**SSIM**	**ClassID**	**#Sample**	**Rank**	**Fnorm**	**SSIM**
0	5923	16	0.0107	0.9962	**T-shirt/top**	6001	17	0.0048	0.9971
1	6742	10	0.0029	0.9992	**1Trouser**	6001	11	0.0001	0.9997
2	6742	15	0.0074	0.9974	**Pullover**	6001	16	0.0027	0.9974
3	6131	15	0.0065	0.9975	**Dress**	6001	14	0.0019	0.9987
4	5842	14	0.0071	0.9976	**Coat**	6001	17	0.0031	0.9969
5	5421	14	0.0097	0.9977	**Sandal**	6001	15	0.0096	0.9953
6	5918	16	0.0068	0.9982	**Shirt**	6001	18	0.0056	0.9950
7	6265	13	0.0090	0.9971	**Sneaker**	6001	14	0.0025	0.9984
8	5841	15	0.0083	0.9987	**Bag**	6001	17	0.0036	0.9988
9	5949	14	0.0056	0.9986	**Ankle boot**	6000	19	0.0020	0.9991

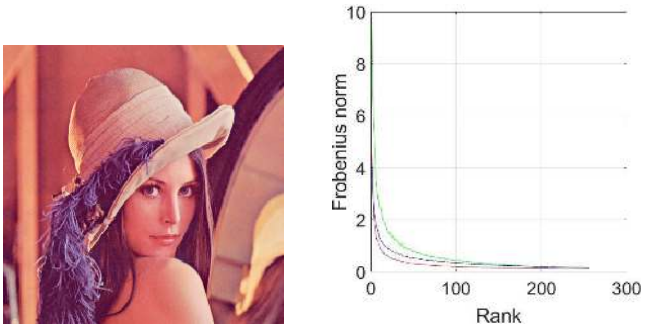
As it evident from Table I that the performance based on the SSIM metric is greater than 0.9950% among all classes in both MNIST data sets, which confirms the validity of our proposed rank truncation method.

B. Approximating the rank for large-size images

In this section, we select miscellaneous standard color images of size 256×256 and 512×512, which published by the Computer Vision Group of the University of Granada[2]. Where the resolution of such images are very high, and approximating their rank for the NMF factorization in their RGB domain is considered substantial [27]; in the view of fact that a lot of information is vulnerable to be lost during images transformation to the grayscale or in resizing their dimensions. Accordingly, it is important to approximate the rank at R-, G-, and B-domain separately to keep an appropriate level of the information at each, then extracting the rank for each domain. Fig. 3(b) highlights the Frobenius norm of the Lena image of size 256×256 and its reconstructed version after applying the β-NMF with $\beta = 1$. Where the rank identified following the traditionally used approach, by varying the rank from the lowest $r = 1$ to highest $r = 256$.

As it can be noticed from the Fig. 3 that the highest Frobenius difference between the original and the reconstructed

[2]http://decsai.ugr.es/cvg/index2.php



(a) (b)

Fig. 3. Approximating the rank for Lena's image. Where (a) Lena image, (b) the Frobenius norm as a function of the rank. The red, green, blue curves in (b) represent the Frobenius norm as a function of rank for R-, G-, B-domain, respectively.

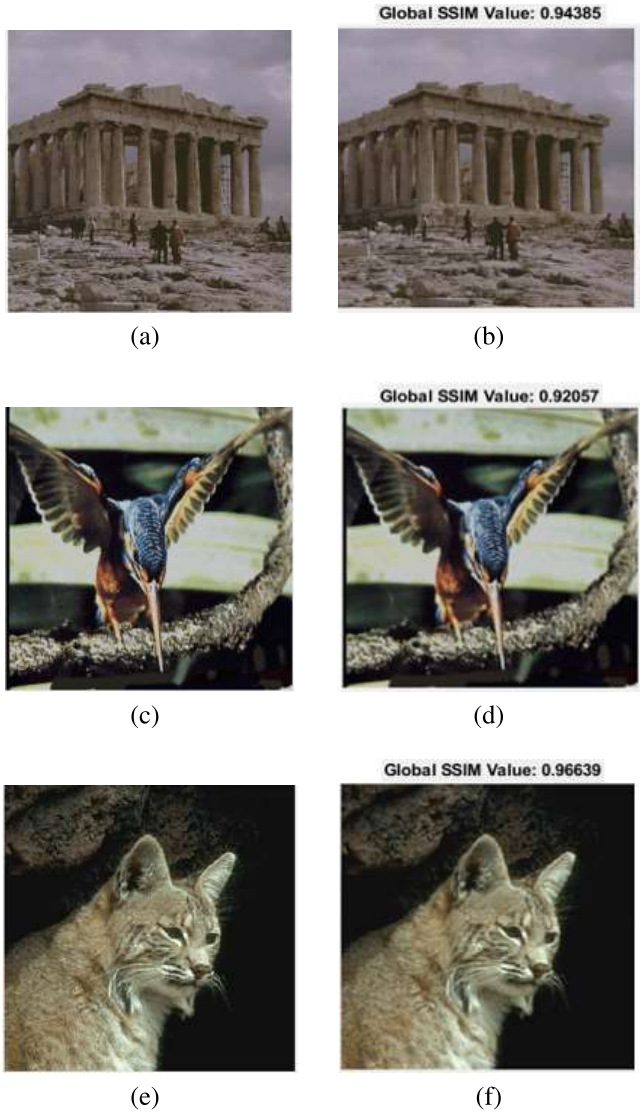
image can be noticed at the first 50 ranks, and the performance is improved as a function of rank around $r = 130 : 170$. The proposed truncation threshold in Eqn. 9 gives $r = 141$ for R domain, $r = 165$ for G domain, and $r = 193$ for B domain for Lena image, which is applicable with the rank plot in Fig. 3(b). Moreover, the SSIM index is equal to 0.9880 according to our truncation threshold, where it is lower than the full rank (when $r = 256$) by 0.0060. Thus, the proposed rank threshold in Eqn. 9 is able to preserve a higher level of the original structure (for large-size images) after the NMF factorization, by retaining the appropriate bounds of singular values that reflect the non-negative rank among the data.

The application domain is extended to a new series of natural images with a size of 512×512 for each one in Fig. 4. By utilizing the same rank truncation proposed in Eqn. 9, where we fixed the evaluation indicator to the SSIM which is commonly used to compare the large-sized images in terms of the structure and the local mapping.

C. Related works comparison

The rank selection stage is considered the first step to carry out the NMF factorization, which initializes the dimensions of both W and H subspaces. Moreover, the rank selection is usually tricky, and there are limited practical approaches available to identify the rank r. For that aim, we proposed a novel method to identify the NMF rank automatically, and we compare our method with the common practical approaches.

Table II reports the performance evaluation of our proposed method employing the same images used in Fig. 1. We compare our results with respect to *(i)* trial and error method by adapting the rank from 1 to full image size [14], *(ii)* retaining 99% of the energy of singular values contribute to the total sum [15], *(iii)* the truncated SVD with keeping 90% of the whole singular values [14], and *(iv)* the MDL [16]. Moreover, we show the average execution time for each method to obtain the rank, also we evaluate the stability when reconstructing the images (the stability is application dependent and we measured the SSIM among the reconstructed and original images, thus the method that gives the highest SSIM is considered the most stable one).



(a) (b) (c) (d) (e) (f)

Fig. 4. Generalizing the rank truncation threshold for natural images with a size of 512×512 for each, where the first column contains the original images and the next one represents the corresponding reconstructed ones. The SSIM equal to 0.9438, 0.9205, and 0.9663 for (b), (d), and (f).

TABLE II
RELATED METHODS COMPARISONS, WHERE THE SAME IMAGES OF FIG. 1 USED FOR THE PERFORMANCE ANALYSIS.

Comparison	Indicators		
Method	Rank	CPU Time	SSIM
Trial and Error [14]	$1 : 28$	NA	$0.4732 : 0.9995$
99% of Energy of Singular Values [15]	10	5.259 s	0.9866
90% of Truncated Singular Values [14]	9	4.781 s	0.9839
MDL [16]	**13**	16.903 m	0.9943
Our Method	**15**	**3.873 s**	**0.9955**

As can be noticed from Table II, that our method achieves the minimum execution time and high reconstruction accuracy based on the SSIM, especially when comparing our result to the recent method (MDL) [16]. Although the MDL achieves similar results to our method, it requires time greater than our method by 262 times to obtain the rank of the set of images used in Fig. 1. Also, the trial and error method that is used in the practice is considered time-consuming and requires

to increase the rank by 1 at each factorization round. Where the mean CPU time needed to factorize each MNIST image using β-NMF is 0.1200 s, provided that the factorization rank is imposed. For the trial and error method, we found that the factorization time when adapting all ranks is equal to $0.1200 \text{ s} \times 28$, and if the image size is relatively large such method is considered time-wasting. Moreover, the truncated singular values [14] however it achieved the lower SSIM, but it requires an average factorization time greater than our method by 0.908 s for the same 100 images. Instead, our proposed method reduces the required time to carry out the factorization by automatically derive the rank from the data itself, and our performance analysis shows superior results in terms of preserving the similar amount of the structure information when using all possible ranks.

Our method is able to approximate the NMF rank with a lower computational time with respect to methods in the literature, at the same time obtaning a higher similarity to the original images. The proposed approach can therefore be used to efficiently map the data to reduced subspaces, with limited information loss. Due to its advantages, it could be used to optimize the learning process of deep ML architectures, by using a compact data representation that can reduce the computational times required to perform the training. Furthermore, our approach could prove beneficial when dealing with devices with limited computational resources (e.g., mobile CPUs or FPGAs), or in methodologies requiring real-time learning (e.g., online learning).

V. CONCLUSIONS

The NMF is a linear factorization technique that enforces non-negativity constraints among the factorized subspaces, which leads to precisely additive positive components to the factorized subspaces that are parts from the original data.Thus, the NMF offers the interpretability when it is integrated into the XAI models, due it its ability to relate the factorized subspaces to the original data. The NMF reduce the dimensionality but it requires identifying and impose the non-negative rank r before carrying out the factorization, that usually unknown in advance. To solve this problem, we proposed a novel rank truncation method that reflects the rank by counting the singular values which lie on certain bounds and consider their number as the rank. The rank is obtained directly according to the truncation threshold, instead of iteratively adapting the rank. Also, it evaluated using β-NMF for the small-sized images of size 28×28 and generalized for the large-sized-images of size 512×512. In future works, we plan to investigate to what extent the learning among the NMF representations (W, H) able to accelerate the deep learning models as in utilizing the autoencoder to encode W, or H, to compress the weights, i.e. partial learning.

REFERENCES

[1] D. Gunning, "Explainable artificial intelligence (xai)," *Defense Advanced Research Projects Agency (DARPA), nd Web*, vol. 2, 2017.

[2] A. Genovese, V. Piuri, and F. Scotti, "Towards explainable face aging with generative adversarial networks," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 3806–3810.

[3] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 2018, pp. 80–89.

[4] A. Coates, A. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 2011, pp. 215–223.

[5] V. V. Vesselinov, M. K. Mudunuru, S. Karra, D. O'Malley, and B. S. Alexandrov, "Unsupervised machine learning based on non-negative tensor factorization for analysis of reactive-transport site data simulations."

[6] A. Genovese, V. Piuri, K. N. Plataniotis, and F. Scotti, "Palmnet: Gabor-pca convolutional networks for touchless palmprint recognition," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 12, pp. 3160–3174, 2019.

[7] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, p. 788, 1999.

[8] B. Alexandrov, V. V. Vesselinov, and H. N. Djidjev, "Non-negative tensor factorization for robust exploratory big-data analytics," Los Alamos National Lab, Los Alamos, NM (United States), Tech. Rep., 2018.

[9] A. N. Langville, C. D. Meyer, R. Albright, J. Cox, and D. Duling, "Initializations for the nonnegative matrix factorization," in *Proceedings of the twelfth ACM SIGKDD international conference on knowledge discovery and data mining*. Citeseer, 2006, pp. 23–26.

[10] P. Magron and T. Virtanen, "Towards complex nonnegative matrix factorization with the beta-divergence," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2018, pp. 156–160.

[11] N. Gillis, L. T. K. Hien, V. Leplat, and V. Y. Tan, "Distributionally robust and multi-objective nonnegative matrix factorization," *arXiv preprint arXiv:1901.10757*, 2019.

[12] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[13] S. M. Atif, S. Qazi, and N. Gillis, "Improved svd-based initialization for nonnegative matrix factorization using low-rank correction," *Pattern Recognition Letters*, vol. 122, pp. 53–59, 2019.

[14] N. Gillis, "The why and how of nonnegative matrix factorization," *Regularization, Optimization, Kernels, and Support Vector Machines*, vol. 12, no. 257, pp. 257–291, 2014.

[15] A. Rajaraman and J. D. Ullman, *Mining of massive datasets*. Cambridge University Press, 2011.

[16] S. Squires, A. Prügel-Bennett, and M. Niranjan, "Rank selection in nonnegative matrix factorization using minimum description length," *Neural computation*, vol. 29, no. 8, pp. 2164–2176, 2017.

[17] R. H. Coding, "Information theory," *Prentice Hall*, 1986.

[18] P. A. Parrilo and S. Khatri, "On cone-invariant linear matrix inequalities," *IEEE Transactions on Automatic Control*, vol. 45, no. 8, 2000.

[19] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM review*, vol. 52, no. 3, pp. 471–501, 2010.

[20] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge univ 2012.

[21] R. Haag, *Local quantum physics: Fields, particles, algebras*. Springer Science & Business Media, 2012.

[22] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM journal on Matrix Analysis and Applications*, vol. 21, no. 4, pp. 1253–1278, 2000.

[23] X. Peng, C. Lu, Z. Yi, and H. Tang, "Connections between nuclear-norm and frobenius-norm-based representations," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 1, pp. 218–224, 2016.

[24] S. Gu, L. Zhang, W. Zuo, and X. Feng, "Weighted nuclear norm minimization with application to image denoising," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014.

[25] L. Deng, "The mnist database of handwritten digit images for machine learning research [best of the web]," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[26] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[27] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.