UNSUPERVISED LEARNING FROM LIMITED AVAILABLE DATA BY β-NMF AND DUAL AUTOENCODER

Mohanad Abukmeil, Stefano Ferrari, Angelo Genovese, Vincenzo Piuri, and Fabio Scotti
Department of Computer Science, Università degli Studi di Milano, Italy
{*firstname.lastname*}@unimi.it

ABSTRACT

Unsupervised Learning (UL) models are a class of Machine Learning (ML) which concerns with reducing dimensionality, data factorization, disentangling and learning the representations among the data. The UL models gain their popularity due to their abilities to learn without any predefined label, and they are able to reduce the noise and redundancy among the data samples. However, generalizing the UL models for different applications including image generation, compression, encoding, and recognition faces different challenges due to limited available data for learning, diversity, and complex dimensions. To overcome such challenges, we propose a partial learning procedure by utilizing the β-Non Negative Matrix Factorization (β-NMF), which maps the data into two complementary subspaces constituting generalized driven priors among the data. Moreover, we employ a dual-shallow Autoencoder (AE) to learn the subspaces separately or jointly for image reconstruction and visualization tasks, where our model performance shows superior results to the literary works when learning the model with a small amount of data and generalizing it for large-scale unseen data.

Index Terms— Unsupervised Learning, Limited data learning, Non-negative Matrix Factorization, Autoencoder.

1. INTRODUCTION

Unsupervised Learning (UL) models have the abilities to operate with unlabeled data to perform dimensionality reduction, and representation learning [1, 2]. The UL models include Principal Component Analysis (PCA) [3], Independent Component Analysis (ICA), Autoencoder (AE), Non-negative Matrix Factorization (NMF), Tensor Decomposition (TD), and others [4]. Among all UL models, the NMF is the only one that decomposes the data into two non-negative subspaces: the first one is termed as W latent space and the other is named as H mixing space [5]. The NMF subspaces constitute rooted priors to learn data, because they hide the positive features, sparse, and part-based representations that represent the original data [6].

Recently, there is a demand to generalize Machine Learning (ML) models for real-life applications, where the data and computational resources are limited or scarce to carry out learning [7, 8, 9]. In most cases, the learning procedure in recent works uses more samples in the training stage than the testing: around 70% of the data are used for the training and the other for the validation and testing [10]. Moreover, different works feed ML models by data without preprocessing and considering learning the relevant representation, thus the models show more bias and overfitting [11, 12, 13].

To overcome the above issues, we propose a novel method to learn with limited available data (fewer samples for training than the testing), utilizing the β-NMF factorization due to its ability in providing a driven prior among the image data and lead to generalize the learned model for large-scale unseen data, *i.e.*, out-of-distribution generalization. The β-NMF maps the data into two positive and sparse subspaces that constitute parts of the original data (rooted representations), thus the representations learning can be facilitated when building shallow ML models to learn among the data [14].

Since any NMF method requires to impose a factorization rank that reflects the original data space dimensions [15] to be learned, we will propose a novel approach to approximate the NMF rank. Moreover, a shallow AE will be employed to learn the factorized subspaces partially (partial AE) or jointly using Dual-shallow AE. The rest of this paper is organized as follows. Section 2 highlights the β-NMF and the shallow AE. Section 3 shows the proposed learning methodology. The experimental results will be given in Section 4. Section 5 reports the conclusion and future works.

2. THE β-NMF AND AE LEARNING

2.1. β-NMF Factorization

The NMF is a linear dimensionality reduction method that belongs to Blind Source Separation (BSS) models, due to its ability to learn part-based representations in an unsupervised way [16]. Also, it is used in image recognition and reconstruction applications: for a given image $X \in \mathbb{R}^{m \times n}$, it decomposes the data matrix as $X \approx WH + R_s$, where $W \in \mathbb{R}^{m \times r}$ represents the bases of the latent subspace, $H \in \mathbb{R}^{r \times n}$ contains the mixing subspace, R_s is the residual, and r is the factorization rank. The NMF represents the data as a product to two subspaces, where the objective function of the optimization procedure minimizes the residual R_s by measuring the mismatch between the original data and the reconstructed subspaces.

The most widely used class of objective function is termed as β-divergence which comprising the Itakura-Saito (IS) when $\beta = 0$, Kulback Leibler (KL) when $\beta = 1$, and Frobenius norm when $\beta = 2$ [14]. The role of such objective functions is to quantify the distance between the original data and the two factored subspaces, *i.e.*, W and H [17]. 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 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 (or data matrix), $\tilde{X} = W_r H_r$, W_r and H_r are the bases of the latent space and the coefficients of the mixing space, respectively, and resulting by the factorization using rank r. To achieve the minimum divergence, the

matrix update procedure [18] is followed as:

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

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

where T is the matrix transpose and $\odot$ denotes the element-wise multiplication. The factorization rank can be estimated by using the Singular Value Decomposition (SVD) to diagonalize the data [19], then considering the number of the singular values larger than a suitable threshold, ϵ. For the experiments described in the following, ϵ has been defined as: $\epsilon = \frac{\sqrt{\mathrm{Tr}(X)}}{||X||_* + ||X||_{\mathrm{F}}}$, where $\mathrm{Tr}(X)$ is the trace, $||X||_*$ is the nuclear norm and is obtained from the SVD, and $||X||_{\mathrm{F}}$ is the Frobenius norm. The threshold reflects a suitable rank approximation for NMF factorization, because it retains a suitable bound of the singular values that constitute the ground truth data.

2.2. AE Learning

AE models are unsupervised generative models that can be formed in shallow or deep architectures, which are utilized in ML to perform dimensionality reduction, recognition, and generation tasks [11]. They also share a similar goal in capturing the hidden structure among the data, by reconstructing through the samples and benefiting from the advantages of the encoding and decoding stages [20]. For a given image (or data sample) $X \in \mathbb{R}^D$ (D is the input image dimensions), the encoding stage provides a mapping $f : \mathbb{R}^D \rightarrow \mathbb{R}^d$, $0 < d < D$ (d is the bottleneck layer dimensions), to corresponding encoded data $Z = f(X; \theta_e)$, while the decoding stage provides a decoding mapping $g : \mathbb{R}^d \rightarrow \mathbb{R}^D$, which reconstructs an approximation of the input data: $\tilde{X} = g(Z; \theta_d)$. Commonly, f and g can be composed of several encoding decoding stages (deep architecture) with a high degree of symmetry, in which the mappings of the j^{th} stage are parameterized by a weight matrix W_j and bias b_j. The objective function minimizes the reconstruction error is given as:

$$F_{(\theta_e,\theta_d)} = \arg\min \|X - (\theta_e \circ \theta_d^{\intercal})X\|^2_{\mathrm{Er}} \quad (5)$$

where $\theta_e = \{W_e, b_e\}$ contains the encoder's weights and biases, $\theta_d = \{W_d, b_d\}$ comprises the decoder's weights and biases, and $\circ$ is the Hadamard product of two matrices that gives element-wise commutative product $(\theta_e \circ \theta_d) = (\theta_d \circ \theta_e)$. Such product gives a realization (in terms of data matrix) that the decoding weights are similar to the transpose of the ones that are used for encoding, or trained to be similar, *i.e.*, $\theta_d \approx \theta_e{}^{\intercal}$. The reconstruction loss $\mathrm{E_r}$ can be measured by different metrics including Mean Square Error (MSE), Frobenius norm, β- divergence, and a recently utilized one for image applications is the Structure Similarity Index (SSIM) [21].

The main challenge in AE learning lies in finding and generalizing both encoding and decoding parameters (θ_e and θ_d), which minimize the reconstruction loss [22]. Especially, when only limited data available for learning, or when the computational resources restrict learning from big data sets (as in high order tensor data) [7]. The β-NMF helps the AE to learn reduced sparse, non-negative, and part-based representations from both W and H subspaces, while feeding the AE with the original data enforces the AE to learn the data with noises, redundancy, and costly due to the original data dimensions. We will show that our approach achieves a minimum reconstruction loss by comparing the related works, specifically, when generalizing the trained model for large-scale unseen samples. Also, we will show how our proposed method allows capturing the rich representations that retain the fidelity among the factorized and encoded data to the manifold of the original data.

3. THE PROPOSED AE LEARNING METHODOLOGY

The NMF factorization has been utilized in building AE models to achieve a large-scale generalization. The Non-negative Sparse AE (NNSAE) proposed in [10] for online learning, enforces the weights of the AE hidden layer to be positive by using an asymmetric regularization and logistic activation function among neurons. Moreover, the NNSAE approach has been followed in [23] to build the Non-negative Constrained AE (NCAE) model for image reconstruction and classification. The recent work proposed the AE with a simplified Random Neural Network (AERNN) [24], where a training rule similar to the NMF used and able to update the model's weights in a non-negative way. However, such models do not exploit the NMF as a data-driven method to disentangle rooted representations and conscious prior among the data. Our proposed methodology is characterized from the others by decomposing the data in an unsupervised way to be in a volume where the noise and redundancy are removed, and then uses the AE to learn the representations. Thus, the learned model is generalized for a large scale unseen data.

The concept of consciousness prior has been proposed for Natural Language Processing (NLP) applications, to combine different priors for disentangling abstract factors to be learned in further stages [25]. Such priors are seen as a bottleneck from which the extracted factors or representations have to proceed for further processing, and lead to generalization improvement for ML and AE models. Following similar footprints of [25], we employ the β-NMF to derive consciousness priors among the data, followed by learning each β-NMF subspace using a shallow AE for each. The subspace W offers a prior among the data's hidden structure, while the H space contains a prior of the mixing coefficients that reconstruct the data.

To show the performance of our proposed work, we employ both MNIST digits and MNIST fashion data sets [26, 27]. Each comprises 60k images for training and 10k for the testing stage, divided in 10 classes with image size of 28×28. However, to challenge the generalization ability of our model to large unseen data only 10k images will be trained, and the testing performance will be measured on the other data set samples. The proposed methodology is divided in the following two steps:

- The initial stage includes image factorization to extract the β-NMF subspaces (W, H), utilizing the rank identification threshold in (ϵ) and setting $\beta = 1$. The rank has been estimated using a subset of the training set (1k images from each class) for each the data sets, to be generalized among the whole data samples. The first and third classes from both data sets have been employed to measure the rank threshold robustness, then the Frobenius norm has been applied to measure the factorization loss acquired by the β-NMF, *i.e.*, the difference between the factorized data and the original one. We obtained $r = 16$ and $r = 15$ (for the first and third class, respectively) from the MNIST digits, and $r = 16$ and $r = 17$ (for the first and third class, respectively) from the MNIST fashion data set. The averaged factorization loss among all testing samples (5k images from each class) did not exceed 0.017. Similarly, the rank is approximated among all classes in both data sets, where the generalized rank (taken from ϵ) $r = 16$ for both data sets due to their image size similarity.
- The second stage is dedicated to the AE learning among the β-NMF subspaces, separately by a shallow AE, or jointly by a dual-shallow AE (Dual-AE) as depicted in Fig. 1. Each AE for either separate or joint learning shares the same number of layers: one layer for each encoder, decoder, and a bottleneck layer. The bottleneck's size still an open problem in the AE

learning, thus we followed [28] approach to identify the required number of neurons, where a discussion about the size of the bottleneck of the different data sets given therein. The method in [28] is implemented by halving the feature vector dimensions and imposing the number as a bottleneck size. In our experiments, the feature dimensions are $28 \times r = 448$, but we expand the bottleneck size for the W-AE (see Fig. 1) to 250 neurons due to its sparsity nature, and we reduce it to 200 neurons to the H-AE (196 neurons proposed in [28]). For the MNIST fashion, because the class complexity, we just expand the size to 400 and 300 for the W-AE and H-AE, respectively. Finally, all AE experiments have been fixed under 2000 epochs, saturated liner encoder and decoder transfer function, $l_2 = 0.0001$ and sparsity regularizer with coefficient $= 0.01$.

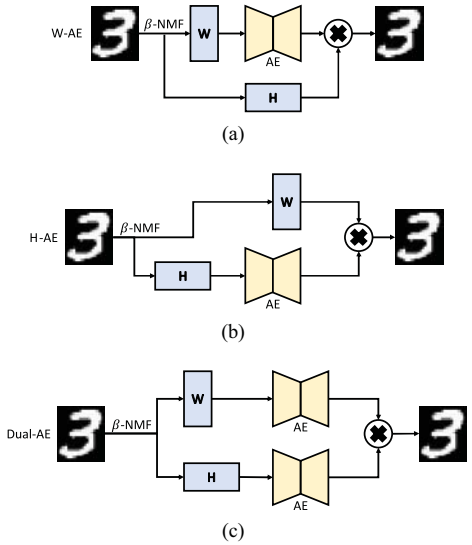


Fig. 1: The three proposed AE schemes, where (a) W-AE in the top, (b) H-AE in the middle, and (c) Dual-AE in the bottom.

4. PERFORMANCE EVALUATION

4.1. Experimental Results

We considered three scenarios indicated as *(i)* W-AE and *(ii)* H-AE when only W and H subspaces are learned, respectively, and *(iii)* Dual-AE when both the subspaces are jointly learned to reconstruct the data, see Fig. 1. We employed both MNIST data sets to compare our method with the literature in terms of MSE error, which is commonly used in shallow AE learning. Besides, to show the ability of our method to preserve the original data structure, we also employed the SSIM index. For more details about the SSIM we refer to [21].

(i) W-AE: The latent space W hides sparse, non-negative and part-based representations, which can be obtained from the β-NMF and learned by the AE to reconstruct the data. The W space contains a lot of sparse values that require a wider bottleneck layer than H-AE ($250, 400$ neurons for the MNIST digits and MNIST fashion, respectively). We measured to which extent that W space can be learned separately, while the H space is used for the reconstruction (Fig.1 (a)). Moreover, the learning complexity can be carried out at $O(m \times r)$ where m is the row space dimension and r is the data rank, instead of learning at $O(m \times n)$ where n is the column space dimension. Table 1 shows the reconstruction performance of the W-AE model.

	Training		Testing	
W-AE	SSIM	MSE	SSIM	MSE
MNIST Digits	0.935	0.005	0.923	0.005
MNIST Fashion	0.760	0.021	0.759	0.022

Table 1: Reconstruction performance of W-AE model.

(ii) H-AE: The mixing space H is considered a combiner to reconstruct the original data, where it is multiplied by W space for the reconstruction purpose. Also, it is less sparse than W space, thus we reduce the number of neurons in the AE bottleneck layer to 200 and 350 for the MNIST digits and MNIST fashion, respectively. As in learning W space, we measured to which extent that H space can be learned, while keeping W space as identity for the sake of reconstruction (Fig. 1b). The learning complexity carried out in $O(r \times n)$, instead of learning at $O(m \times n)$. Table 2 shows the reconstruction performance of the H-AE based on the used indicators in Table 1.

	Training		Testing	
H-AE	SSIM	MSE	SSIM	MSE
MNIST Digits	0.986	0.004	0.961	0.005
MNIST Fashion	0.888	0.006	0.872	0.007

Table 2: Reconstruction performance of H-AE model.

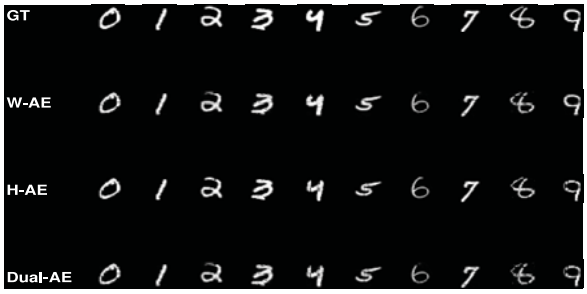
(iii) Dual-AE: In this scenario both W and H spaces are learned separately to reconstruct the data jointly utilizing a dual AE: the model is fully automated and avoids keeping W or H as identities (as in the W-AE or H-AE) for the reconstruction. Moreover, the learning complexity for the Dual-AE scenario is $O(m \times r + r \times n)$. Accordingly, the learning process can be facilitated and implemented among different machines; we carried out the learning for the Dual-AE in two separated machines with core $i7$-CPU at each one, where the learning complexity was $O(m \times r)$ and $O(r \times m)$ for the first and second CPU, respectively. Table 3 highlights the reconstruction performance of the Dual-AE model.

	Training		Testing	
Dual-AE	SSIM	MSE	SSIM	MSE
MNIST Digits	0.910	0.009	0.900	0.010
MNIST Fashion	0.778	0.017	0.767	0.018

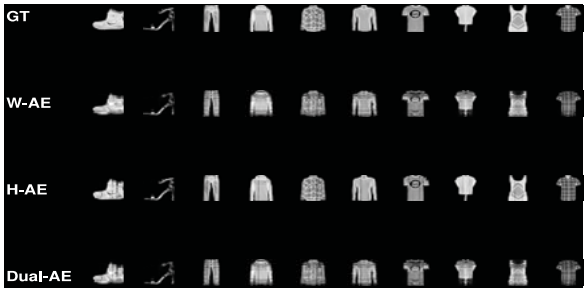
Table 3: Reconstruction performance of Dual-AE model.

To show the reconstruction ability of the W-AE, H-AE, and Dual-AE, Fig. 2 depicts the reconstruction differences between each AE with respect to the ground truth data. As it can be noticed from Fig. 2 and the above tables, that the H-AE outperforms the others AE in terms of data reconstruction.

Eventually, to show the fidelity of the factorized and the encoded data to the original one, in terms clustering and preserving the inter-classes variations, Fig. 3 shows the t-SNE [29] embedding of the original MNIST digits data set, the factorized space W of the β-NMF, and the latent (bottleneck) space of the W-AE model. As it can be concluded from Fig. 3, that the factorized space W and the AE's latent space maintain the discriminating features to be clustered, and both show the same fidelity to the original data set; preserving the clustering properties and avoiding classes' shuffling (*i.e.*, realizing an invariant transform). Finally, the computational complexity of the t-SNE is reduced from $O(DN^2)$ (Fig. 3a) where D is the dimensions of each image (for the MNIST samples $D = 28 \times 28$) and N is the number of samples in the data set to $O(d_r N^2)$ (Fig. 3b) where d_r is the W space dimensions $d_r = 28 \times 16$, See Section 3.

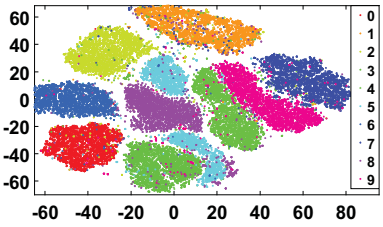


(a) MNIST Digits Reconstruction

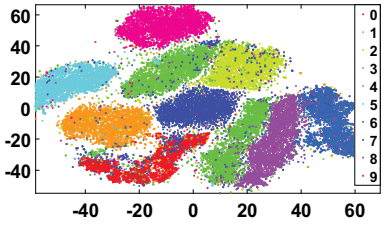


(b) MNIST Fashion Reconstruction

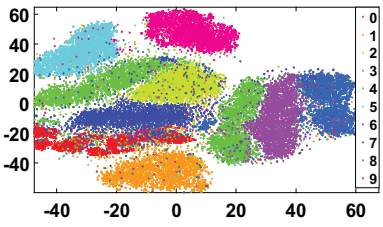
Fig. 2: The reconstruction of MNIST data sets (testing samples) where GT represents the Ground Truth samples, W-AE represents the W-AE reconstruction, H-AE represents the H-AE reconstruction, Dual-AE represents the reconstruction of dual AE learning.



(a) Original Data Embedding



(b) W Space Embedding



(c) AE Latent Space Embedding

Fig. 3: The t-SNE of the original MNIST digit data set, W space of the factorized data set, and the latent (bottleneck) space of the W-AE. 40000 samples have been employed for each sub-fig.

4.2. Recent works comparison

The performance comparison of the basic recent deep UL models based on the NMF and AE learning is reported in Table 4. It also comprises: Stacked AE with Restricted Boltzmann Machine (SAE-RBM) [30], Non-Negative Sparse AE (NNSAE) [10], fold-AE [31], Group Sparse AE (GSAE) [32], AE Spiking Neural Networks (AE-SNN) [28], Structuring AE (SAE) [33], and Non-Negative AE with Simplified Random Neural Network (NNAE-sRNN) [24]. Every work except [10] and [31] employed 60k training samples and 10k testing to carry out the learning and performance evaluation. In [10] only 10k training samples were used and 50k considered as testing samples, and in [31] only a mini data set of 200 images were used. We followed [10] and employed 10k samples for the training stage and the other for the testing. To show a fair comparison, we adjusted the size of the data set in the range $[200, 10000]$ with 70% training size and the other for the testing, to demonstrate how the testing loss can be saturated. As it can be noticed from Fig. 4, the testing loss shows a minimum around a data set size of 1000, but it is saturated around a data size of $8000 - 10000$, which meets the specifications of our training size (See section 3). Finally, our method outperforms the others in terms of obtaining the minimum reconstruction loss.

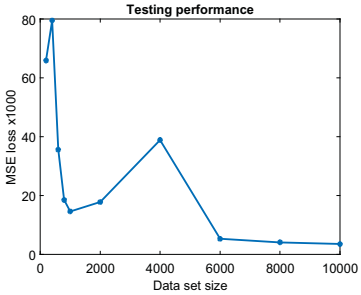


Fig. 4: The Testing loss performance as a function of data set size.

Method	Training MSE		Testing MSE	
	MNIST		MNIST	
	Digits	Fashion	Digits	Fashion
SAE-RBM [30]	0.823	NA	NA	NA
NNSAE [10]	0.012	NA	0.015	NA
Fold-AE in [31]	0.178	NA	5.929	NA
Group Sparse-AE [32]	1.10	1.10	NA	NA
AE-SNN [28]	0.110	0.150	0.122	0.178
Structuring-AE [33]	0.025	0.014	NA	NA
NNAE-sRRNN [24]	0.024	NA	NA	NA
Our proposed W-AE	**0.005**	**0.021**	**0.005**	**0.022**
Our proposed H-AE	**0.004**	**0.006**	**0.005**	**0.007**
Our proposed Dual-AE	**0.009**	**0.017**	**0.010**	**0.018**

Table 4: Comparison with recent methods in the literature, including deep unsupervised learning methods [32, 28, 33, 24].

5. CONCLUSIONS

We proposed an approach based on unsupervised data factorization and encoding, to be utilized for ML tasks as in image reconstruction and visualization. We used the β-NMF to reduce the data dimensionality and obtain both latent and mixing spaces W and H, respectively. We trained a shallow AE at each space, and used a dual-shallow AE to learn from both spaces jointly. The performance analysis shows that our proposed work obtained the minimum reconstruction loss when it is compared with the relevant works, especially when learning the model with limited available data (a small set for training but large for testing). In future works, we plan to investigate our proposed work for other ML tasks such as regression or classification. Also, we plan to extend the dual shallow AE to learn from tensor data with order > 2 (multi-way) as in the RGB or hyper-spectral image learning, where we need n-way AE.

6. REFERENCES

[1] A. Coates, A. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *Proc. of the 14th Int. Conf. on Artificial Intelligence and Statistics*, 2011, pp. 215–223.

[2] H. B. Barlow, "Unsupervised learning," *Neural computation*, vol. 1, no. 3, pp. 295–311, 1989.

[3] H. Elaydi, M. Alhanjouri, and M. Abukmeil, "Palmprint recognition using 2d wavelet, ridgelet, curvelet and contourlet," *i-manager's Journal on Electrical Engineering (JEE)*, vol. 7, no. 1, pp. 9–19, 2013.

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

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

[6] A. Cichocki, R. Zdunek, A. H. Phan, and S.-i. Amari, *Non-negative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*, John Wiley & Sons, 2009.

[7] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan, "Deep learning with limited numerical precision," in *Int. Conf. on Machine Learning*, 2015, pp. 1737–1746.

[8] Y. Cai, A. Genovese, V. Piuri, F. Scotti, and M. Siegel, "Iot-based architectures for sensing and local data processing in ambient intelligence: Research and industrial trends," in *2019 IEEE Int. Instrumentation and Measurement Technology Conference (I2MTC)*. IEEE, 2019, pp. 1–6.

[9] M. Abukmeil, S. Ferrari, A. Genovese, V. Piuri, and F. Scotti, "On approximating the non-negative rank: Applications to image reduction," in *Proc of the 2020 IEEE Int. Conf. on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA 2020)*, Tunis, Tunisia, June 2020.

[10] A. Lemme, R. F. Reinhart, and J. J. Steil, "Efficient online learning of a non-negative sparse autoencoder," in *Proc. of ESANN*, 2010.

[11] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE TPAMI*, vol. 35, no. 8, pp. 1798–1828, 2013.

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

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

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

[15] 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.

[16] Y.-X. Wang and Y.-J. Zhang, "Nonnegative matrix factorization: A comprehensive review," *IEEE Trans on Knowledge and Data Engineering*, vol. 25, no. 6, pp. 1336–1353, 2012.

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

[18] 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.

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

[20] P. Baldi, "Autoencoders, unsupervised learning, and deep architectures," in *Proc. of the ICML Workshop on Unsupervised and Transfer Learning*, 2012, pp. 37–49.

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

[22] M. Längkvist, L. Karlsson, and A. Loutfi, "A review of unsupervised feature learning and deep learning for time-series modeling," *Pattern Recognition Letters*, vol. 42, pp. 11–24, 2014.

[23] E. Hosseini-Asl, J. M. Zurada, and O. Nasraoui, "Deep learning of part-based representation of data using sparse autoencoders with nonnegativity constraints," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 27, no. 12, 2015.

[24] Y. Yin and E. Gelenbe, "Non-negative autoencoder with simplified random neural network," in *Proc. of the 2019 Int. Joint Conf. on Neural Networks (IJCNN)*, July 2019, pp. 1–6.

[25] Y. Bengio, "The consciousness prior," *arXiv preprint arXiv:1709.08568*, 2017.

[26] 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.

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

[28] D. Roy, P. Panda, and K. Roy, "Synthesizing images from spatio-temporal representations using spike-based backpropagation," *Frontiers in Neuroscience*, vol. 13, pp. 621, 2019.

[29] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. November, pp. 2579–2605, 2008.

[30] C. C. Tan and C. Eswaran, "Reconstruction and recognition of face and digit images using autoencoders," *Neural Computing and Applications*, vol. 19, no. 7, pp. 1069–1079, 2010.

[31] J. Wang, H. He, and D. V. Prokhorov, "A folded neural network autoencoder for dimensionality reduction," *Procedia Computer Science*, vol. 13, pp. 120–127, 2012.

[32] A. Sankaran, M. Vatsa, R. Singh, and A. Majumdar, "Group sparse autoencoder," *Image and Vision Computing*, vol. 60, pp. 64–74, 2017.

[33] M. Rudolph, B. Wandt, and B. Rosenhahn, "Structuring autoencoders," in *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV) Workshops*, October 2019.