ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

Non-Negative Tensor Dictionary Learning

Abraham Traoré, Maxime Berar and Alain Rakotomamonjy *

LITIS, Normandie Université, University of Rouen
76800 Saint-Etienne du Rouvray, FRANCE
abraham.traore@etu.univ-rouen.fr
maxime.berar,alain.rakotomamonjy@univ-rouen.fr

Abstract. A challenge faced by dictionary learning and non-negative matrix factorization is to efficiently model, in a context of feature learning, temporal patterns for data presenting sequential (two-dimensional) structure such as spectrograms. In this paper, we address this issue through tensor factorization. For this purpose, we make clear the connection between dictionary learning and tensor factorization when several examples are available. From this connection, we derive a novel (supervised) learning problem which induces emergence of temporal patterns in the learned dictionary. Obtained features are compared in a classification framework with those obtained by NMF and achieve promising results.

1 Introduction

Dictionary learning has been a key technique in a wide variety of applications in computer vision and signal processing for learning sparse representations of input data. Typical applications are image denoising [1] or signal classification [2]. When data at hand have specific non-negative structures that are important to preserve, dictionary learning leads to the so-called non-negative matrix factorization (NMF) problem whose objective is to decompose a matrix $\mathbf{S}$ into $\mathbf{W}$ and $\mathbf{H}$, with minimal divergence between $\mathbf{S}$ and $\mathbf{WH}$. Typical divergence measures are the Euclidean, the Kullback-Leibler and the Itakura-Saito ones [3].

One of the application domain of NMF is audio analysis in which signals are frequently represented as spectrogram *i.e.* a matrix of time-frequency energy. A major drawback of the widely used Euclidean NMF in this context is that the temporal structure of the TFR is not properly handled. Indeed, such a NMF considers column vectors of the matrix to decompose independently of each other. Several approaches have been considered in order to alleviate such a drawback. For instance, convolutive NMF can discover temporal patterns that occur in the TFR. However, that method is also known to poorly handle variability of patterns [4]. Another approach consists in stacking temporally adjacent frames of $\mathbf{S}$ in order to build a set of vectors capturing large-scale temporal information and then in applying NMF on these stacked vectors. While frame stacking poses problem related to algorithmic complexity due to matrix size augmentation, it sometimes helps achieving very good results for audio signal classification [5]. Reformulating NMF as a tensor decomposition has also

*This work benefited from the support of the project OATMIL ANR-17-CE23-0012 of the French National Research Agency (ANR), the Normandie Projet GRR-DAISI, European funding FEDER DAISI.

been suggested, and discarded without strong justification by Van Hamme as a solution for capturing (temporal) structure in matrix factorization [4]. Our objective in this paper, is to restore faith in tensor factorization for dictionary learning with temporal structure. While tensor-based dictionary learning has already been investigated [1, 6], our work extends the current state-of-the-art by allowing overcompleteness and by leveraging supervised information such as labels in the factorization.

Our contributions are the following: (i) we clarify the connection between dictionary learning, NMF and tensor factorization in a context of supervised representation learning for classification. (ii) We provide insights on why tensor-based dictionary learning TDL is able to learn temporal structures and propose an algorithm for solving the TDL problem. (iii) Experimental results on toy and real-world datasets show that the features learned by our TDL are competitive with those obtained from NMF followed by pooling.

2 Learning Tensor-based Dictionaries

2.1 Tensors and multilinear algebra basics

A tensor is a multidimensional array defined on the tensor product of vector spaces, the number of those spaces being the order. A N-order tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times .. \times I_N}$ (typed in boldface Euler script letters) contains elements denoted by $\mathcal{X}_{i_1,..,i_N}$, $\{I_n\}_{1 \leq n \leq N}$ being the dimensions of the vector spaces whose product defines $\mathcal{X}$. Matricization reorders these elements into a matrix by rearranging the tensor fibers, fibers being defined by fixing every index but one. The mode-n matricization of $\mathcal{X}$ arranges the mode-n fibers to be the columns of the resulting matrix $\mathbf{X}^{(n)} \in \mathbb{R}^{I_n \times (\prod_{k \neq n} I_k)}$. The mode-n product of $\mathcal{X}$ with a matrix $\mathbf{B} \in \mathbb{R}^{J_n \times I_n}$ denoted $\mathcal{X} \times_n \mathbf{B}$ yields a tensor $\mathcal{Y} \in \mathbb{R}^{I_1 \times .. \times J_n \times .. \times I_N}$ defined by: $\mathbf{Y}^{(n)} = \mathbf{B}\mathbf{X}^{(n)}$. Analogous to the matrix Frobenius norm, we can define a tensor Frobenius norm by: $\|\mathcal{X}\|_F^2 = \sum_{i_1,..,i_N} \mathcal{X}_{i_1,..,i_N}^2$. This analogy can also be applied to the ℓ_1 norm $\|\mathcal{X}\|_1 = \sum_{i_1,..,i_N} |\mathcal{X}_{i_1,..,i_N}|$.

The two most commonly used decompositions of tensors are the Tucker and the Parafac decompositions [7]. Given $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$, classical Tucker decomposition looks for the approximation: $\mathcal{X} \approx \mathcal{G} \times_1 \mathbf{A}^{(1)} \times .... \times_N \mathbf{A}^{(N)}$, with $\mathcal{G} \in \mathbb{R}^{J_1 \times ..... \times J_N}$ being the core tensor and the matrices $\mathbf{A}^{(n)} \in \mathbb{R}^{I_n \times J_n}$ the loading factors with $J_n \leq I_n$(because the general purpose of Tucker is data compression). The canonical (a.k.a. Parafac) decomposition can be seen as a special case of Tucker decomposition with $J_i = J_j, \forall i, j$ and $\mathcal{G}$ diagonal.

2.2 From NMF to Tensor decomposition

Given a set of L signals represented as spectrograms $\{\mathbf{S}_i\}_{i=1}^{L} \in \mathbb{R}_+^{F \times T}$ and an integer K, fixing the number of dictionary elements, a classical formulation of the dictionary learning via NMF is the following optimization problem:

$$\min_{\mathbf{W}\in\mathbb{R}_+^{F\times K},\mathbf{H}_i\in\mathbb{R}_+^{K\times T}} \sum_{i=1}^{L} \|\mathbf{S}_i - \mathbf{W}\mathbf{H}_i\|_F^2 + \lambda_1\|\mathbf{W}\|_F^2 + \lambda\sum_{i=1}^{L}\|\mathbf{H}_i\|_1. \qquad (1)$$

The matrix $\mathbf{W} \in \mathbb{R}_+^{F\times K}$ contains the dictionary elements as columns and such dictionary is overcomplete if $K \geq F$. The second term downweighs each dictionary atom norm and the third one is used to make the activation matrices $\mathbf{H}_i \in \mathbb{R}_+^{K\times T}$ sparse (thus enabling to learn in an overcomplete setting)

Let's now consider a third order tensor $\mathcal{S} \in \mathbb{R}_+^{L\times F\times T}$ whose horizontal slices $\{\mathcal{S}_{i,:,:}\}_{i=1}^{L}$ are the spectrograms $\{\mathbf{S}_i\}_{i=1}^{L}$. The problem (1) can be rewritten as:

$$\min_{\mathbf{W}\in\mathbb{R}_+^{F\times K},\mathcal{H}\in\mathbb{R}_+^{L\times K\times T}} \|\mathcal{S} - \mathcal{H}\times_2\mathbf{W}\|_F^2 + \lambda_1\|\mathbf{W}\|_F^2 + \lambda\|\mathcal{H}\|_1, \qquad (2)$$

with the horizontal slices of $\mathcal{H}$ being the matrices $\{\mathbf{H}_i\}_{i=1}^{L}$. The equivalence between the two problems arises from mode-2 matricization ($\|\mathbf{S}^{(2)} - \mathbf{W}\mathbf{H}^{(2)}\|_F^2$). As matricization depends on an ordering convention, it highlights that any set of dictionary atoms $\{\mathbf{W}_{:,k}\}_{k=1}^{K}$ is invariant to a shuffle of the columns of $\mathbf{S}_i$, *i.e.* any temporal information contained in the samples are ignored in the dictionary.

In order to code temporal patterns of frequency atoms shared by the samples, for example onset/offset or piecewise-constant temporal patterns, one should introduce a matrix $\mathbf{W}^{(t)}$ containing K_t temporal dictionary atoms to complement the K_f frequency atoms of $\mathbf{W}^{(f)}$. The model should approximate each spectrogram as a linear combination of the $K_f \times K_t$ basis elements $\left\{\mathbf{W}_{:,p}^{(f)}(\mathbf{W}_{:,q}^{(t)})^\top\right\}$, allowing overcompleteness when $F \times T \leq K_f \times K_t$. This leads to the following optimization problem for TDL:

$$\min_{\mathbf{W}^{(f)},\mathbf{W}^{(t)},\mathcal{H}} \|\mathcal{S} - \mathcal{H}\times_2\mathbf{W}^{(f)}\times_3\mathbf{W}^{(t)}\|_F^2 + \lambda_1^f\|\mathbf{W}^{(f)}\|_F^2 + \lambda_1^t\|\mathbf{W}^{(t)}\|_F^2 + \lambda\|\mathcal{H}\|_1$$
$$\text{s.t} \quad \mathbf{W}^{(f)} \in \mathbb{R}_+^{F\times K_f}, \mathbf{W}^{(t)} \in \mathbb{R}_+^{T\times K_t}, \mathcal{H} \in \mathbb{R}_+^{L\times K_f\times K_t}. \qquad (3)$$

This is the so-called Tucker-2 decomposition [7], with penalty terms inherited from the dictionary learning problem. The intuition behind this learning problem is that owing to the norm residual minimization and non-negativity constraints, temporal structures of high-energy shared across samples will emerge from $\mathbf{W}^{(t)}$. The disposal of $\mathbf{W}^{(f)}$ and $\mathbf{W}^{(t)}$ is justified by the definition of $\mathcal{S}$, whose 2^{nd} and 3^{rd} modes represent frequency and time. It is also worth to notice that the tensor $\mathcal{H}$ size increases with K_f and K_t.

2.3 Classification framework

As our main objective is to classify a set of spectograms according to a set of labels, we exploit supervision in the tensor factorization problem by adding a novel loss term :

$$\min_{\mathbf{W}^{(f)},\mathbf{W}^{(t)},\mathcal{H},\mathbf{B}^{(f)},\mathbf{B}^{(t)}} \|\mathcal{S} - \mathcal{H}\times_2\mathbf{W}^{(f)}\times_3\mathbf{W}^{(t)}\|_F^2 + \lambda_1^f\|\mathbf{W}^{(f)}\|_F^2 + \lambda_1^t\|\mathbf{W}^{(t)}\|_F^2 \qquad (4)$$

$$+\lambda\|\mathcal{H}\|_1 + \lambda_c\|\mathcal{C} - \mathcal{H} \times_2 \mathbf{B}^{(f)} \times_3 \mathbf{B}^{(t)}\|_F^2 + \lambda_2^f\|\mathbf{B}^{(f)}\|_F^2 + \lambda_2^t\|\mathbf{B}^{(t)}\|_F^2,$$

$$\text{s.t} \quad \mathbf{W}^{(f)} \in \mathbb{R}_+^{F\times K_f}, \mathbf{W}^{(t)} \in \mathbb{R}_+^{T\times K_t}, \mathcal{H} \in \mathbb{R}_+^{L\times K_f\times K_t}, \mathbf{B}^{(f)} \in \mathbb{R}_+^{K_f\times K_f}, \mathbf{B}^{(t)} \in \mathbb{R}_+^{K_t\times K_t},$$

As in [2], the role of $\mathbf{B}^{(f)}$ and $\mathbf{B}^{(t)}$ is to make the activation coefficients aligned with the label information brought by the tensor $\mathcal{C}$ built similarly to the matricial formulation presented in [2]. Numerous dictionary learning algorithms based on tensor decomposition have already been presented [1, 6]. However, no proposed approaches can be applied straightforwardly to our problem formalization. The TDL problem as presented above is solved by alternate minimization of the convex problems resulting from fixing all factors but one. The update of $\mathcal{H}$ is done via the resolution of a non-negative least squares. Since the update of $\mathbf{W}^{(f)}, \mathbf{W}^{(t)}, \mathbf{B}^{(f)}, \mathbf{B}^{(t)}$ can be large-scale problems (for K_f, K_t large), we have chosen a projected gradient method, which is known for its efficiency in solving large scale convex minimization problems subject to linear constraints [8]. After the dictionaries inference, we recompute activation coefficients by projecting each spectrogram on the obtained dictionaries independently of the class information regularization and feed their vectorization to a classifier.

3 Numerical experiments

These experiments aim at illustrating that our tensor-based approach is able to learn temporal structures in time-frequency representations, while a typical NMF followed by a global pooling may fail in capturing these structures. As such, we have compared the features obtained from TDL and from *supervised* (label information is used) NMF [2] with *max* and *average* pooling in multi-class classification problems. For both approaches, the number of spectral atoms in the decompositions is fixed to the same value and resulting feature vectors are fed to a SVM classifier with Gaussian kernel. For all problems, hyperparameters, including SVM ones, have been selected through cross-validation on the training set. The initialization is performed for the NMF problem by drawing uniform numbers on $[0, 1]$ and for TDL, by solving NMF of the matricized forms of $\mathcal{S}$ and $\mathcal{C}$. Details for reproducibility are available upon request.

Synthetic data set: This data set is composed of 400 examples equally split into 8 classes. Each signal is a sum of two localized sinusoids of different temporal scales on which a uniform noise is added. We transform each signal into a spectrogram, resized (for computational reasons) into a non-negative matrix of size 25 × 25. Examples of TFR templates for all classes are given in Figure 1. The performance of all the algorithms have been evaluated through their mean average precision. Learning curves of

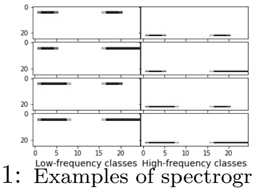

Fig. 1: Examples of spectrogram template for each of the class. On the left and right, we respectively have low-frequency and high-frequency signals with different temporal scales.

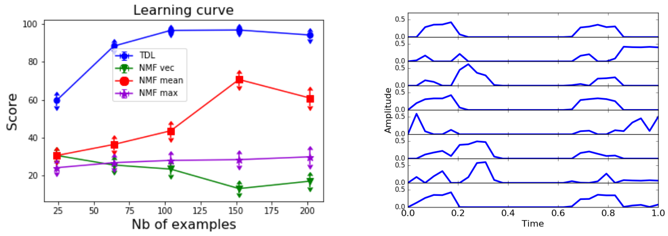

Fig. 2: (left) Learning curves of all algorithms. NMF vec, mean and max respectively refer to features derived by performing vectorization, mean pooling and max pooling operation on activation matrix associated to NMF. (right) Example of temporal dictionary atoms showing that TDL is able to learn temporal structures.

all algorithms are given in Figure 2, where a point is the average performance over 5 trials (with different noise realizations). Our tensor-based approach is able to produce discriminative features with only 24 examples (3 per classes) and it outperforms all NMF approaches regardless of the pooling strategy. On the right plot of Figure 2, we have depicted 8 elements of temporal dictionary $\mathbf{W}^{(t)}$. We can note that each of these atoms represents temporal scale of the sinusoids composing the signals to be classified. More importantly, they display different temporal lengths showing that the learned temporal dictionaries are robust to the signal scales.

DCase2013 data set: This dataset is for acoustic scene detection problem. However, audio scenes in this dataset do not have temporal structures suited to TDL (since discriminative sound events can occur at any moment in the 30-s scene). Our goal in this experiment is to prove empirically that TDL is still able to learn relevant temporal dictionary atoms. The dataset is composed of ten scene categories recorded in different locations. Training and test sets are made of 100 30-second-long scene instances with ten examples per class [9]. Spectrograms for these signals are represented as non-negative matrices of size 60×20. Again, all hyperparameters have been tuned by cross-validation. Figure 3 reports the performance of NMF with pooling and our TDL for increasing number of spectral dictionary atoms and 10 temporal atoms. We can note that best performance of all methods are nearly equal with slight advantage for NMF+max pooling. Interestingly, max pooling also helps in stabilizing performance. More interestingly, the right plot of the Figure 3 shows that as there are few temporal structures to be learned by tensor decomposition, TDL learns temporal dictionary atoms that behave as localized mean pooling covering all the time span.

4 Conclusion

In this paper, we have developed a tensor-based overcomplete dictionary learning framework able to infer frequent temporal patterns in sequential data. The novel framework proposed makes clear the connection between dictionary learning and

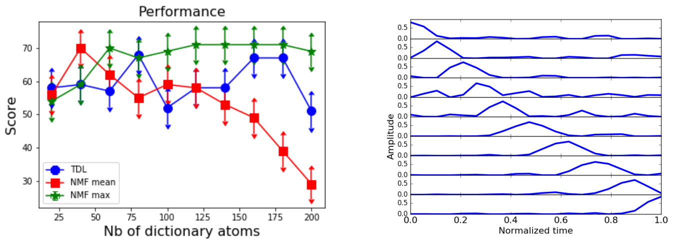

Fig. 3: (left) Performance of TDL and NMF + pooling on DCase 2013. (right) Example of temporal dictionary atoms.

tensor factorization and can be easily extended with supervised information. Our experimental results show that the learned features are competitive compared to those obtained using NMF followed by pooling. The main bottleneck of our approach is computational and we plan to lift this issue by exploring an online version of our algorithm.

References

[1] Andrzej Cichocki and Anh Huy Phan. Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *IEICE transactions on fundamentals of ECCS*, 92(3):708–721, 2009.

[2] A Rakotomamonjy. Supervised Representation Learning for Audio Scene Classification . *IEEE/ACM Transactions on ASLP*, 25:1253–1265, 2017.

[3] Cédric Févotte and Jérôme Idier. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural computation*, 23(9):2421–2456, 2011.

[4] Hugo Van Hamme. An On-Line NMF Model for Temporal Pattern Learning: Theory with Application to Automatic Speech Recognition. *LVA/ICA. LNCS*, 7191:306–313, 2012.

[5] Justin Salamon and Juan Pablo Bello. Unsupervised feature learning for urban sound classification. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 171–175, 2015.

[6] Florian Roemer, Giovanni Del Galdo, and Martin Haardt. Tensor-based algorithms for learning multidimensional separable dictionaries. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3963–3967, 2014.

[7] Tamara G. Kolda and Brett W. Bader. Tensor Decompositions and Applications. *SIAM REVIEW*, 51(3):455–500, 2009.

[8] Rafal Zdunek and Andrzej Cichocki. Fast nonnegative matrix factorization algorithms using projected gradient approaches for large-scale problems. *Intell. Neuroscience*, 2008:3:1–3:13, 2008.

[9] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley. Detection and classification of acoustic scenes and events. 17(10):1733–1746, 2015.