ORTHOGONALITY-REGULARIZED MASKED NMF WITH KL-DIVERGENCE FOR LEARNING ON WEAKLY LABELED AUDIO DATA

Iwona Sobieraj, Lucas Rencker, Mark D. Plumbley

University of Surrey
Centre for Vision Speech and Signal Processing
Guildford, Surrey GU2 7XH, United Kingdom

ABSTRACT

Non-negative Matrix Factorization (NMF) is a well established tool for audio analysis. However, it is not well suited for learning on weakly labeled data, i.e. data where the exact timestamp of the sound of interest is not known. To overcome this shortcoming of NMF, we recently proposed the Orthogonality-Regularized Masked NMF (ORM-NMF), that allows to extract meaningful representations from weakly labeled audio data. Here we extend the method to allow it to use the generalized Kullback-Leibler (KL) divergence as a cost function of NMF. We demonstrate that the proposed Orthogonality-Regularized Masked NMF with KL divergence can be used for Audio Event Detection of rare events and evaluate the method on the development data from Task2 of DCASE2017 Challenge.

1. INTRODUCTION

In the recent paper, inspired by the idea of forcing two dictionaries to be different originally introduced for single channel source separation problem [1], we proposed to add an orthogonality regularizer, that decorrelates the two dictionaries and promotes orthogonality between dictionaries of the sound and of the background, resulting in the Orthogonality - Regularized Masked NMF (ORM-NMF) method [2]. It achieved promising results on detection of gunshot sounds. The method used Euclidean distance to measure the reconstruction error in the training of NMF. Here, we extend the method by using the KL divergence as a cost function of NMF, which has shown to be more suitable for audio applications.

2. NON-NEGATIVE MATRIX FACTORIZATION

The goal of NMF is to approximate a non-negative data matrix, typically a time-frequency representation of a given sound, $\mathbf{V} \in {\mathbb{R}_+}^{F \times T}$ as a product of a dictionary $\mathbf{W} \in {\mathbb{R}_+}^{F \times K}$ and its activation matrix $\mathbf{H} \in {\mathbb{R}_+}^{K \times T}$, such that:

$$\mathbf{V} \approx \hat{\mathbf{V}} = \mathbf{WH}. \tag{1}$$

The research leading to these results has received funding from the European Union's H2020 Framework Programme (H2020-MSCA-ITN-2014) under grant agreement n 642685 MacSeNet. MDP is also partly supported by EPSRC grant EP/N014111/1

$\mathbf{W}$ and $\mathbf{H}$ are estimated to minimize some divergence metric $D(\mathbf{V}|\mathbf{WH})$. For any two matrices X and Y, we define $D(\mathbf{X}|\mathbf{Y}) = \sum_{m,n} D(x_{mn}, y_{mn})$. In this work we choose the generalized Kullback-Leibler (KL) as the divergence metric, defined as

$$D_{KL}(\mathbf{V}|\mathbf{WH}) = \sum_{k,l} \left(\mathbf{V}_{k,l} \log \frac{\mathbf{V}_{k,l}}{(\mathbf{WH})_{k,l}} - \mathbf{V}_{k,l} + (\mathbf{WH})_{k,l} \right). \tag{2}$$

3. ORM-NMF WITH KL DIVERGENCE

In [3] we proposed to extend a standard NMF approach to learning on weakly labeled data. To explain the idea, let us consider the task of detection of rare sound events, as proposed in the Detection and Classification of Acoustic Scenes and Events challenge DCASE2017 [4]. Let $y \in \{0, 1\}$ be a weak label denoting absence or presence of the target sound, $\mathbf{V}^0 = \mathbf{V}^0_1, \cdots, \mathbf{V}^0_{M_0}$ is a set of M_0 training examples with absence of the target sound and $\mathbf{V}^1 = \mathbf{V}^1_1, \cdots, \mathbf{V}^1_{M_1}$ is a set of M_1 training examples with the presence of the target sound. As the data is weakly labeled, examples containing the target sound most probably also contain noise and other sounds. Therefore, we assume that to reconstruct well the target sound training examples ($\mathbf{V}^1$) we also need elements from dictionaries extracted from background sounds examples ($\mathbf{V}^0$). At the same time, we do not expect elements of the dictionary atoms of target sounds to be used for reconstructing $\mathbf{V}^0$. We impose this constraint in the training phase by applying a binary mask to the activation matrix as follows:

$$\begin{aligned} \mathbf{V} = [\mathbf{V}_0, \mathbf{V}_1] &\approx [\mathbf{W}_0, \mathbf{W}_1] \left(\begin{bmatrix} \mathbf{1} & \mathbf{1} \\ \mathbf{0} & \mathbf{1} \end{bmatrix} \odot \begin{bmatrix} \mathbf{H}_{00} & \mathbf{H}_{01} \\ \mathbf{H}_{10} & \mathbf{H}_{11} \end{bmatrix} \right) \\ &= [\mathbf{W}_0, \mathbf{W}_1] \begin{bmatrix} \mathbf{H}_{00} & \mathbf{H}_{01} \\ \mathbf{0} & \mathbf{H}_{11} \end{bmatrix} \\ &= [\mathbf{W}_0\mathbf{H}_{00}, \mathbf{W}_0\mathbf{H}_{01} + \mathbf{W}_1\mathbf{H}_{11}] \end{aligned} \tag{3}$$

where $\mathbf{W}^0 \in {\mathbb{R}_+}^{F \times K^0}$, $\mathbf{W}^1 \in {\mathbb{R}_+}^{F \times K^1}$ are "sound" and "background" dictionaries respectively, K^0 and K^1 are their corresponding ranks. $\mathbf{0}$ is a matrix of zeros with K^1 rows and the number of columns corresponding to the total size of M_0 background training data, while $\mathbf{1}$ denotes matrices

of appropriate dimensions with all elements equal to 1. $\mathbf{H}^{00}$, $\mathbf{H}^{01}$, $\mathbf{H}^{10}$ and $\mathbf{H}^{11}$ are parts of the activation matrix of suitable dimensions. We then further improve the separation of the dictionaries by adding an additional orthogonality regularizer, which minimizes the coherence between the dictionaries. Combining the constraint on the activation matrix and the orthogonality regularizer results in the following cost function to minimize:

$$\begin{aligned}&\min_{\mathbf{W}_0,\mathbf{W}_1,\mathbf{H}\geq 0} D_{KL}(\mathbf{V}|\mathbf{W}\mathbf{H}) + \lambda\|\mathbf{W}_1^\intercal\mathbf{W}_0\|^2 \\ &= D_{KL}(\mathbf{V}_0|\mathbf{W}_0\mathbf{H}_{00}) + D_{KL}(\mathbf{V}_1|(\mathbf{W}_0\mathbf{H}_{01} + \mathbf{W}_1\mathbf{H}_{11})) \\ &+ \lambda\|\mathbf{W}_1^\intercal\mathbf{W}_0\|^2\end{aligned} \tag{4}$$

As $\|\mathbf{W}_1^\intercal\mathbf{W}_0\|^2$ is convex in $\mathbf{W}_0$ and $\mathbf{W}_1$, we can minimize the cost function using the gradient decent. Then, following the derivations of Lee and Sung [5], we obtain the corresponding multiplicative update rules for $\mathbf{W}_0$ and $\mathbf{W}_1$:

$$\begin{aligned}\mathbf{W}_0 &\leftarrow \mathbf{W}_0 \odot \frac{\frac{\mathbf{V}_0\mathbf{H}_{00}{}^\intercal}{\mathbf{W}_0\mathbf{H}_{00}} + \frac{\mathbf{V}_1\mathbf{H}_{01}{}^\intercal}{\mathbf{W}_0\mathbf{H}_{01}+\mathbf{W}_1\mathbf{H}_{11}}}{\mathbf{1}\cdot\mathbf{H}_{00}{}^\intercal + \mathbf{1}\cdot\mathbf{H}_{01}{}^\intercal + \lambda\mathbf{W}_1\mathbf{W}_1{}^\intercal\mathbf{W}_0} \\ \mathbf{W}_1 &\leftarrow \mathbf{W}_1 \odot \frac{\frac{\mathbf{V}_1\mathbf{H}_{11}{}^\intercal}{\mathbf{W}_0\mathbf{H}_{01}+\mathbf{W}_1\mathbf{H}_{11}}}{\mathbf{1}\cdot\mathbf{H}_{11}{}^\intercal + \lambda\mathbf{W}_0\mathbf{W}_0{}^\intercal\mathbf{W}_1}\end{aligned} \tag{5}$$

As the regularizer does not influence the activation matrix $\mathbf{H}$, the update rule for $\mathbf{H}$ remains the same as in the original NMF problem formulation:

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^T\frac{\mathbf{V}}{\mathbf{W}\mathbf{H}}}{\mathbf{W}^T\mathbf{1}}. \tag{6}$$

Final dictionaries $\mathbf{W}_0$ and $\mathbf{W}_1$ are then concatenated to form a dictionary $\mathbf{W}$ used for audio event detection, whereas the activation matrix $\mathbf{H}$ is discarded.

4. EXPERIMENTAL RESULTS

The proposed method is evaluated on gunshot detection using only weakly labeled data from the audio recordings of the TUT Rare Sound Events 2017. The dataset was provided for Task 2 of the DCASE2017 challenge [4]. We resample the data to 16000 Hz. We extract mel-spectrograms with 40 components, using a window size and hop size of 64 ms. In order to model temporal dynamics we group 4 consecutive frames into 2D patches. We train the system on audio chunks of 4 seconds and evaluate on recordings of 30 seconds. We set the number of positive and negative atoms as $K_0 = 20$, $K_1 = 10$, values chosen empirically.

4.1. Evaluation metrics

To evaluate the method we use event-based error rate (ER) and event-based F-score. An event is considered correctly detected using onset-only condition with a collar of 500 ms. The ER is calculated by adding the number of insertions and deletions for each class before dividing it by the total number of events. The F-score is based on the total amount of false negatives, true positives and false positives [6].

Table 1: Evaluation results of gunshot detection. Error Rate (ER) and F-score (F1) are reported for the Euclidean distance and KL-divergence with different weight of the regularization (λ). When λ = 0, the method becomes standard Masked NMF.

λ	Euclidean		KL-divergence	
	ER	F1	ER	F1
0	1.20	0.50	1.16	0.43
500	1.02	0.56	1.08	0.48
1000	0.92	0.59	**0.80**	**0.64**
5000	1.0	0.56	0.96	0.51

5. CONCLUSION

In this work we showed that the idea of Orthogonality- Regularized Masked NMF can be extended for KL-divergence as the measure of the reconstruction error. We show that ORM-NMF with KL-divergence performs well for audio event detection using weakly labeled data. Using NMF we end up with a parsimonious model comprising only of 30 dictionary atoms, a size which much smaller than contemporary deep learning models. In future work, we plan to generalize the method for other divergences.

6. REFERENCES

[1] E. M. Grais and H. Erdogan, "Discriminative nonnegative dictionary learning using cross-coherence penalties for single channel source separation.," in *Proc. of INTERSPPECH 2013*, p. 808 812.

[2] I. Sobieraj, L. Rencker, and M. D. Plumbley, "Orthogonality-regularized masked NMF for learning on weakly labeled audio data," in *Proc. of ICASSP 2018*.

[3] I. Sobieraj, Q. Kong, and M. D. Plumbley, "Masked Non-negative Matrix Factorization for bird detection using weakly labeled data," in *Proc. of EUSIPCO 2017*.

[4] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proc. of DCASE 2017*.

[5] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization.," *Nature*, vol. 401, no. 6755, pp. 788–91, 1999.

[6] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.