2017 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 25–28, 2017, TOKYO, JAPAN

A NEURAL NETWORK ALTERNATIVE TO CONVOLUTIVE AUDIO MODELS FOR SOURCE SEPARATION

Author(s) Name(s) omitted for double blind review

Author Affiliation(s) omitted for double blind review

ABSTRACT

Index Terms— Auto-encoders, source separation, deep learning.

1. INTRODUCTION

Recently, several neural network architectures have been proposed to develop algorithms for supervised source separation and speech enhancement. (Chandna monaural, Grais Single channel, Park Fully convolutional, Venkataramani end-to-end source). Currently, these networks are trained to learn discriminative audio models extensively. In other words, the spectrogram of the mixture is given as an input to the network. The goal of the network then, is to learn suitable time-frequency masks that separate the input spectrogram into the source and the interference componenents. Thus, the networks learn a basis decomposition that often works only for a specific source-interference pair, i.e., these models are not transferable. If the interfering signal changes, these networks have to be re-trained to learn suitable models to separate the new interfering signal in the mixture from the source.

A popular technique to learn generative audio models for supervised source separation is the use of Non-negative matrix factorization (NMF). Non-negative matrix factorization (NMF) matrix of non-negative elements $\mathbf{X} \in \mathbb{R}^{\geq \curlyvee}_{\mathbb{M}\times\mathbb{N}}$ as a product of the basis matrix $\mathbf{W}$ and the activation matrix $\mathbf{H}$. The notation $\mathbb{R}^{\geq 0}_{M\times N}$ represents the st of matrices of non-negative elements of size $M \times N$. In this factorization, the basis matrix $\mathbf{W} \in \mathbb{R}^{\geq 0}_{M\times r}$, the activation matrix $\mathbf{H} \in \mathbb{R}^{\geq 0}_{r\times N}$ and r represents the rank of the decomposition. In the case of audio signals, we apply such a factorization on audio spectrograms. In this setting, the columns of $\mathbf{W}$ begin to act as representative basis vectors for the source. The rows of $\mathbf{H}$ indicate the activity of these basis vectors in time.
As shown in [](ICASSP Paper reference...), the notion of non-negative audio modeling can be easily generalized by interpreting it as a neural network. We can interpret NMF as a non-negative auto-encoder in the following manner,

$$1^{\text{st}} \text{ layer: } \mathbf{H} = g(\mathbf{W}^{\ddagger} \cdot \mathbf{X}) \quad (1)$$

$$2^{\text{nd}} \text{ layer: } \mathbf{X} = g(\mathbf{W} \cdot \mathbf{H}) \quad (2)$$

Here, $\mathbf{X}$ represents the input spectrogram, $\mathbf{W}^{\ddagger}$ represents a form of pseudo-inverse of $\mathbf{W}$ and $g(.) : \mathbf{R} \rightarrow \mathbf{R}^{+}$ is an element-wise function that maps a real number to the space of positive real numbers. As before, the columns of $\mathbf{W}$ act as representative basis vectors and the corresponding rows of $\mathbf{H}$ indicate their respective activations. Although non-negativity of the models is not explicitly guaranteed in this formulation, applying a suitable sparsity constraint allows the network to learn suitable non-negative models. Additionally, this interpretation enabled a pathway to propose multi-layer extensions by exploiting the wealth of available neural net architectures that could potentially lead to superior separation performance.

Spectrograms of speech and audio signals incorporate temporal dependencies that span multiple time frames. However, NMF and its neural net equivalent are unable to explicitly utilize these cross-frame patterns available in a spectrogram. To alleviate this drawback, Smaragdis [](Paris convolutive speech bases) proposed a convolutive version to NMF that allows spectro-temporal patterns as representative basis elements. In this paper, we develop a neural network alternative to such convolutive audio models for supervised source separation. In doing so, we solve two fundamental problems associated with this task. The first step is to develop a suitable neural network architecture to learn convolutive audio models in an adaptive manner. Utilizing the models to separate a source from a given mixture forms the second step. The remainder of the paper is organized as follows. In section **??**, we develop an auto-encoder that can act as an equivalent to conv-NMF audio models. Section **??** proposes a novel approach to utilize these models for supervised source separation. We evaluate these models in terms of their separation performance in section **??** and conclude in section **??**.

Thanks to XYZ agency for funding.

978-1-5090-6341-3/17/$31.00

2. NON-NEGATIVE CONVOLUTIIONAL AUTO-ENCODERS

2.1. Network Architecture

The convolutive NMF model [](Smaragdis convolutive speech bases) approximates a non-negative matrix $\mathbf{V} \in \mathbb{R}^{\geq 0}_{M \times N}$ as,

$$\mathbf{V}(f,t) \approx \sum_{i=1}^{r} \sum_{k=0}^{T-1} \mathbf{W}_i(k,f) \cdot \mathbf{H}(i,t-k) \tag{3}$$

Here, $\mathbf{W}_i \in \mathbb{R}^{\geq 0}_{M \times T}$ acts as the i^{th} basis matrix and $\mathbf{H} \in \mathbb{R}^{\geq 0}_{r \times N}$ contains the corresponding weights. The notation $\mathbf{V}(i,j)$ represents the element of $\mathbf{V}$ indexed by the i^{th} row and the j^{th} column. Thus, we can interpret this operation as a two-step convolutional auto-encoder as follows,

$$1^{\text{st}} \text{ layer: } \mathbf{H}(i,t) = \sum_{i=1}^{r} \sum_{j} \sum_{k=0}^{T-1} \mathbf{W}_i^{\ddagger}(j,k) \mathbf{X}(j,t-k) \tag{4}$$

$$2^{\text{nd}} \text{ layer: } \mathbf{V}(f,t) = \sum_{i=1}^{r} \sum_{k=0}^{T-1} \mathbf{W}_i(k,f) \cdot \mathbf{H}(i,t-k) \tag{5}$$

Here, we assume that the i convolutional layer filters $\mathbf{W}$, $\mathbf{W}^{\ddagger}$ have a size of $M \times T$ where, T represents the depth of the convolution.

The approximation $\hat{X}$ for a given spectrogram X is computed as follows:

$$\begin{aligned} \hat{H}(k,t) =& \sigma_1 \left(\sum_{f,t'} X(f,t-t') F_e(f,t',k) \right) \\ \hat{X}(f,t) =& \sigma_2 \left(\sum_{k} \sum_{t'} \hat{H}(k,t-t') F_d(f,t',k) \right) \end{aligned} \tag{6}$$

3. EXTENSIONS TO INFINITE SUMMATION

This is the same as CNN-CNN case, except the computation of the activations H. In the CNN encoder, each filter was of finite length. With RNN-CNN version, we are attempting to use an infinite length filter. The computation of $\hat{H}$ is as follows:

$$\begin{aligned} Z(:,k,t) =& \sigma(WZ(:,k,t-1) + UX(:,t)) \\ \hat{H}(k,t) =& \sum_{f} Z(f,k,t) \end{aligned} \tag{7}$$

Give a toy example which shows what this model can do that CNN-CNN can not.

4. SUPERVISED SOURCE SEPARATION

Describe the separation procedure.

5. EXPERIMENTS

Try each model in a given K range in speech-speech source separation task.

5.1. Experimental setup

5.2. Results and Discussion

6. CONCLUSION