A Neural Network Alternative to Non-Negative Audio Models

PARIS SMARAGDIS#*
SHRIKANT VENKATARAMANI#

#UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
*ADOBE RESEARCH

ICASSP 2017
Introduction

- **Supervised Single-channel Source Separation**
  - Given a mixture of $N$ sources
    \[ x(t) = \sum_{i} w_i s_i(t), \text{ where } w_i \in \mathbb{R} \text{ for } i = \{1, \ldots, N\} \]
  - Separate individual sources
  - Training data in the form of alternate unmixed recordings of the source.

- **Non-negative Matrix Factorization (NMF)**

- **Objective:** Develop a neural network alternative to NMF
Outline

- Non-negative Matrix Factorization (NMF)
- Non-negative Auto-encoder (NAE) equivalent to NMF
- Supervised source separation using NAE models
- Results
NMF

- NMF for matrices

\[ X = WH \quad \text{where} \quad X \in \mathbb{R}_{\geq 0}^{m \times n}, \ W \in \mathbb{R}_{\geq 0}^{m \times r}, \ H \in \mathbb{R}_{\geq 0}^{r \times n}, \]

- NMF is posed as a minimization problem

\[
\begin{align*}
&\text{minimize} \\
&\quad D(X, WH) \\
&\text{subject to} \\
&\quad W \geq 0, \ H \geq 0.
\end{align*}
\]

where \( \geq 0 \) implies element-wise non-negativity

- Commonly used Cost functions
NMF: Piano example

- No cross-cancellations
- Part based decomposition
- Meaningful basis vectors.
- Can be used as a model for supervised source separation.

\[ D = KL( X \parallel WH ) \]
Towards an NMF neural network

- Autoencoder: Reconstructs the input at the output
  - Encoder: Input to Code
  - Decoder: Code to approximation of input

\[ X = WH \quad \text{s.t.} \quad H \geq 0 \]

\[ \hat{X} = WH \quad \text{s.t.} \quad W \geq 0 \]
Towards an NMF neural network

- **Autoencoder**: Reconstructs the input at the output
  - **Encoder**: Input to Code
  - **Decoder**: Code to approximation of input

- $g(x) = \max(x, 0)$ or $\ln(1 + \exp(x))$ or $|x|$ mapping to the space of positive real nos.
Piano Example

Without sparsity

\[ D = K L( X \parallel g(WH) ) \]

With Sparsity

\[ D = K L( X \parallel g(WH) ) + \| H \|_1 \]
Why is this a good idea?

- Allows for several extensions over regular NMF

Recurrent NAE-NMF

Multi-layer NAE-NMF
Supervised source separation

- Learn representative bases for all the sources.
  - Autoencoder training on unmixed training examples gives representative matrices $W_s$ and $W_n$.

- The spectrogram of the mixture is the sum of spectrograms of the sources.

$$X_m = S + N = g(W_s H_s) + g(W_n H_n)$$

Thus,

$$X_m^T = g(H_s^T W_s^T) + g(H_n^T W_n^T)$$

An output neural network with inputs: $W_s^T, W_n^T$ and output: $X_m^T$
Supervised source separation

- Solve the minimization problem for $H_s$ and $H_n$
  \[
  \min_{H_s, H_n} KL( X_m \mid\mid g(W_s H_s) + g(W_n H_n) )
  \]
  Solved by training the output neural network

- Reconstruct the sources
  \[
  \hat{s}_i[n] = \text{STFT}^{-1}\left( \frac{g(W_i H_i)}{\sum_{i \in \{s, n\}} g(W_i H_i)} \otimes X_m \otimes e^{j \Phi_m} \right) \quad \text{for } i \in \{s, n\}
  \]
  where $\Phi_m$ represents the phase of the mixture
  $\text{STFT}^{-1}$ represents the overlap and add STFT operation
Results

NMF vs shallow- NAE
Rank = 20

NMF vs multilayer NAE
Rank = 20
Results

NMF vs shallow-NAE
Rank = 100

NMF vs multilayer NAE
Rank = 100
Conclusion

❑ Non-negative Auto-encoder (NAE) audio models equivalent to NMF
  ❑ Easily generalizable

❑ Separation Performance
  ❑ Shallow NAE models equivalent to NMF
  ❑ Multilayer NAE models outperform NMF by ~ 1.5 dB (SDR)

❑ Future work
  ❑ Alternate neural net architectures for NAE
  ❑ Towards an end-to-end neural net for source separation.
THANK YOU
## Demo

<table>
<thead>
<tr>
<th></th>
<th>Ground truth (SDR = 6.05 dB)</th>
<th>NMF (SDR = 5.4 dB)</th>
<th>Two layer NN (SDR = 7.1 dB)</th>
<th>Four layer NN (SDR = 7.1 dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Source 1</strong></td>
<td><img src="image1" alt="Piano" /></td>
<td><img src="image2" alt="Piano" /></td>
<td><img src="image3" alt="Piano" /></td>
<td><img src="image4" alt="Piano" /></td>
</tr>
<tr>
<td><strong>(Male)</strong></td>
<td><img src="image1" alt="Piano" /></td>
<td><img src="image2" alt="Piano" /></td>
<td><img src="image3" alt="Piano" /></td>
<td><img src="image4" alt="Piano" /></td>
</tr>
<tr>
<td><strong>Source 2</strong></td>
<td><img src="image1" alt="Piano" /></td>
<td><img src="image2" alt="Piano" /></td>
<td><img src="image3" alt="Piano" /></td>
<td><img src="image4" alt="Piano" /></td>
</tr>
<tr>
<td><strong>(Female)</strong></td>
<td><img src="image1" alt="Piano" /></td>
<td><img src="image2" alt="Piano" /></td>
<td><img src="image3" alt="Piano" /></td>
<td><img src="image4" alt="Piano" /></td>
</tr>
</tbody>
</table>