Sound Source Separation via Deep Neural Network

Yichangle Zhao (zhaoyic@stanford.edu)

Introduction

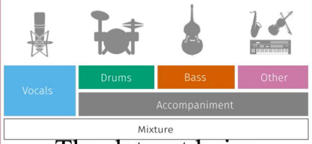
It is an interesting but also challenging task for audience to identify the sound of violin and the sound of viola while listening to a great symphony.

As always, researchers are attracted to conduct researches in this area to see if algorithms can do any better on this task.

First, sound source separation via signal processing was an active research area, but the results were not appealing enough. One of the reasons behind is that even the sounds produced by same kind of instrument can have plenty of variations due to the variations of individual instruments.

Therefore, sound source separation via deep neural networks start to become popular as deep learning has the ability of identifying both differences and similarities.

Dataset

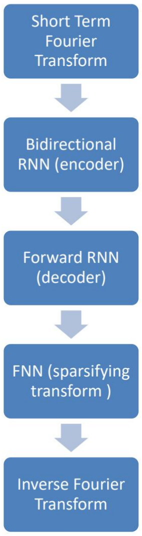


The dataset being employed is Demixing Secrets Dataset 100 (DSD100) from SiSEC. DSD100 contains 100 mixture tracks of different genres along with their sound sources tracks. The sound sources include vocals, drums, bass, and other accompaniment, while the music genres include rap, pop, rock, country, heavy metal, electronic, jazz and reggae. The diversity ensures the neural network model the ability of handling various cases. The dataset is split into a development set and a test set, where the dev set consists of 90 mixture tracks and their sound sources tracks, and the test set consists of the remaining 10 combinations.

Method

- Short term Fourier transform is applied to convert sound tracks into time-frequency representations
- Bidirectional RNN as a encoder, inputs vector representation of sound tracks, outputs its hidden state
- Forward RNN works as a decoder, inputs hidden state from bidirectional RNN, outputs its own hidden state
- L2 loss and L2 regularization are used
- During testing, mixture sound track is converted to time-frequency representation, and predicted filter is applied on top of it, then the result is converted back to sound track via inverse Fourier transform
- Performance evaluated using BSS, especially SDR and SIR

Result



Short Term Fourier Transform

Bidirectional RNN (encoder)

Forward RNN (decoder)

FNN (sparsifying transform)

Inverse Fourier Transform

Two comparables:
GRA3: DNN based supervised learning to predict filter
CHA: CNN based approach to produce estimates of all source signals using an ideal ratio mask (IRM)

	SDR	SIR
GRA3	-1.74	1.28
CHA	1.58	5.17
Masker	1.67	4.14

Future

Possible improvements:

- Train on more data to obtain the ability of handling variety
- Add a denoising layer to have consistent performance
- Add TwinNet architecture as regularizer to take into consideration long-term temporal patterns

References:
Deep neural network based instrument extraction from music (S. Uhlich, F. Giron, and Y. Mitsufuji, 2015)
A recurrent encoder-decoder approach with skip-filtering connections for monaural singing voice separation (S.-I. Mimilakis, K. Drossos, G. Schuller, and T. Virtanen, 2017)
MaD TwinNet: Masker-Denoiser Architecture with Twin Networks for Monaural Sound Source Separation (Konstantinos Drossos, Stylianos Ioannis Mimilakis, Dmitriy Serdyuk, Gerald Schuller, Tuomas Virtanen, Yoshua Bengio, 2018)
Monaural singing voice separation with skip-filtering connections and recurrent inference of time-frequency mask (S.-I. Mimilakis, K. Drossos, J.-F. Santos, G. Schuller, T. Virtanen, and Y. Bengio, 2018)
Twin Networks: Matching the future for sequence generation (D. Serdyuk, N.-R. Ke, A. Sordoni, A. Trischler, C. Pal, and Y. Bengio, 2018)