Sound Source Separation via Deep Neural Network

Yichangle Zhao
zhaoyic@stanford.edu

Abstract

This report is dedicated to a project of sound source separation via deep neural network. Specifically, we focus on separating sound signals of different musical instruments from sound signal mixtures. The report will briefly introduce the problem and its significance, and explain the dataset being used in detail before elaborating the model and techniques get employed. The report will conclude with the discussion of results and possible future improvements.

1. 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. In general, humans are intuitively good at filtering out one or a few individual sounds from a noisy environment. For example, one should be able to realize his or her name is being called even in a party setting where music is loud, and everyone is taking. However, the task of separating sounds of different musical instruments from music mixtures is a bit different: even humans always have trouble doing the task as the sounds of different musical instruments can be very similar to each other and the melody each instrument is playing can be very similar as well.

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. Different attempts have been made, and some researches showed convincing improvements on performance metrics.

This report is dedicated to a deep learning model which can be considered as a filter that is similar to time-frequency masking being used in non-deep-learning solutions. When the filter is applied on the time-frequency representation of a piece of sound mixture, the outputs will be the time-frequency representations of the sounds from different musical instruments. From there, the sounds of different musical instruments can be reproduced, and performance can be evaluated using quantitative metrics. The deep-learning-based model is able to predict the best filter on the task and therefore achieve better results than non-deep-learning solutions.

2. 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. During training process, the 90 samples are furthermore sliced into smaller pieces so that more training samples can be used to obtain better performance. For individual mixture tracks and sound source tracks, the audio channels are stereo, the sample rate is 44.1 kHz, and the number of bits per sample is 16. The sound tracks are later turned into vector representations for further processing.

3. Method

Once the data is ready, the short-term Fourier transform is first applied to convert the sound tracks into time-frequency representations, on which later the filter can be used. And then a frequency trimming step is invoked to reduce the dimension of the input before we feed the data into neural networks, so that the number of

trainable parameters is largely reduced, and hence the computational cost and training time are largely saved as well, without impacting the performance too much.

The model is to predict a time-frequency filter that will be applied to the input mixture magnitude spectrogram, which is obtained from above preprocessing steps, to estimate the magnitude spectrogram of target sound source. In other words, the filter works just like a convolution, which is applied to the vector representation of input mixture and outputs the vector representation of a sound source. Therefore, the model mainly consists of three parts: the first part is a bidirectional recurrent neural network, which works as an encoder, and outputs its hidden state. Next, a forward RNN is employed and works as a decoder on the hidden state of the first bidirectional RNN, then the forward RNN decoder outputs its own hidden state. The hidden state of the forward RNN is then fed into a sparsifying transformation to obtain the time-frequency mask that is aimed for. During the whole process, L2 loss function with L2 regularization is used as the objective of the model.

During testing time, the mixture sound track is first converted to a time-frequency vector representation via short-term Fourier transformation and then the predicted time-frequency mask is applied on top of it. The resulting vector representation of a single sound source later goes through an inverse Fourier transformation, so that the corresponding sound track is generated. And blind source separation (BSS) eval, including signal-to-distortion ratio (SDR) and signal-to-interference ratio (SIR), is used to evaluate the performance of the model.

Although there are three parts of the model, the object is shared and easy to calculate, therefore training the model is not too computationally costly.

4. Results

The performance of the model is evaluated using blind source separation (BSS) eval, in particular, signal-to-distortion ratio (SDR) and signal-to-interference ratio (SIR). The median of SDR is 1.67 while the median of SIR is 4.14.

If we want to compare the results with other well-established algorithms, one of the good candidates is GRA3, which is a DNN based supervised approach to predict the mask that will be used to process the mixture magnitude spectrogram. The median of SDR for GRA3 is -1.74, while the median of SIR for GRA3 is 1.28. Intuitively, our model is doing better than GRA3 in both aspects.

Another good candidate to compare is CHA, which is a CNN approach to produce estimates of all source signals using an ideal ratio mask (IRM). The median of SDR for CHA is 1.58 while the median of SIR for CHA is 5.17. Our model has slightly higher SDR median.

As RNN is always a good fit for processing sequential data, we would expect this model to have even better performance if trained with more data and well-tuned hyperparameters.

5. Discussion

There are still some improvements than can be made on this model to help the performance. If more training data is obtained, the model will be more robust to handle variety and noise. And it is also feasible to add a denoising layer so that the performance can be more consistent.

Another major improvement can be made is adding TwinNet architecture. This is essentially an additional backward RNN which is used as a regularizer for the forward RNN in the model, so that the forward RNN not only focuses on the most recent information, but also takes into consideration long-term temporal patterns. This design has been proved to be an effective way to improve the performance of the neural network, but we leave out this part due to time and resource limitations.

6. Reference

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)