Overview

Combining different models is a common strategy to build a good single channel audio source separation (SCSS) system. We combine fully convolutional neural networks (FCNs) and recurrent neural networks (LSTMs/BLSTMs). FCNs are good at extracting useful features from the audio data and BLSTMs are good at modeling the temporal structure of the audio signals. Our experimental results show that combining FCNs and BLSTMs achieves better separation performance than using each model individually.

Problem formulation of SCSS

Given a mixture of audio sources as \( y(t) = \sum_{i=1}^{N} s_i(t) \), we wish to find estimates \( \hat{s}_i(t) \) of the source audio signals \( s_i(t) \). This can be formulated in the time-domain as \( y(n) = \sum_{i=1}^{N} s_i(n) \), where \( s_i(n) \) is the unknown STFT of source audio \( s_i(t) \).

FCN for source separation

The fully convolutional neural network (FCN) consists of an encoder part and a decoder part. The encoder part consists of repetitions of convolutional layers and an activation layer. Each convolutional layer consists of filters that extract features from the input layer, the activation layer imposes non-linearity to the extracted features. The decoder part consists of repetitions of deconvolutional (transposed convolutional) layer and an activation layer.

The FCN is used to map the magnitude spectrogram of the input mixture into the magnitude spectrogram of the target source. The FCN in this work is a 2D convolutional deep neural network. The input and output data for the FCN are 2D signals (magnitude spectrograms) and the filtering is a 2D operator.

BLSTMs and LSTMs for source separation

The Bidirectional Long Short-Term Memory (BLSTM) is a Long Short-Term Memory (LSTM) recurrent neural network that uses contextual information from the past and future of its input/output sequence. Figure 2 shows the recurrent neural network structure that we use in this work. The hidden layers are BLSTM layers and the output layer is an LSTM layer. The input of the BLSTM is a sequence of N consecutive frames from the spectrograms of the mixed signal. The output of the LSTM layer is a spectral mask corresponding to the N source audio signals.

The FCN is used first to extract an initial magnitude spectrogram of the target source from the input mixture. The initial estimate is then passed to the BLSTM network to enhance the output sequence of the FCN. The FCN is good at extracting useful features from the input signal and the BLSTM is good at modeling the temporal structure of the input sequence.

Joint training for the combined models

The trained models of the FCN and BLSTM models for source \( i \) are stacked to form the combination of the FCN and BLSTM models. FCN-BLSTM. A joint training is then run over the combined model (FCN-BLSTM) to refine the parameters of the trained models for the training data set. The input of the combined FCN-BLSTM model during training is the magnitude spectrogram of the mixed signal and the reference output is the reference spectral mask \( M_{\text{ref}}(f, m) \). The output of the combined FCN-BLSTM model is the output spectrogram \( M_{\text{est}}(f, m) \) computed from Eq. (4). The training of the FCN-BLSTM model is done by minimizing the cost function in Eq. (3).

Experimental

We apply our proposed SCSS using FCN-BLSTM to separate the singing voice from a group of songs from the S2SE-2015-MUS-TASK dataset. We compared the performance of the combined FCN-BLSTM model with each of the individual model (FCN and BLSTM) individually and also with the feedforwards neural network (FFN).

The tables show the number of layers, the type of each layer, the number of filters, units in each layer, the size of the filters, and the total number of parameters for the FCN, BLSTM, FCN-BLSTM and FFN models. The activation function in the FCN and BLSTM layers is the rectified linear unit (ReLU). The activation function in the FFN is sigmoid in the forward direction and hard-sigmoid in the backward direction.

The results indicate that combining the FCN and BLSTM models achieves the best performance of the FCN in SIR (more separation) and the best performance of the BLSTM in SAR (less artifacts). The proposed method of using FCN followed by BLSTM (FCN-BLSTM) works better than BLSTM, even with fewer parameters in the FCN-BLSTM than the BLSTM.

Acknowledgements

This work is supported by grant EP/L027119/2 from the UK Engineering and Physical Sciences Research Council (EPSRC).