Noise Reduction Using Self-Attention Deep Neural Networks

Naoyuki Shiba and Hiroaki Saito

Graduate School of Scienece and Technology, Keio University, Japan

Abstract. In recent years, there has been a lot of research on the task of source separation, which is the separation of sound sources from a piece of music into vocal and accompaniment components. This paper proposes a model that introduces self-attention to Open-Unmix, an open source software for the source separation task. Self-attention is a mechanism that learns and determines the data flow itself in neural networks. We applied this model to a speech separation task to remove noise from speech, and compared it with previous methods using an objective evaluation measures (SDR, ISR, SAR). The results show that a proposed method outperform the previous methods in SDR. Furthermore, the t-test showed a significant difference between the two methods.

1 Introduction

Recently developed deep learning techniques have been used in the study of sound source separation, and their accuracy has been dramatically improved. Source separation refers to the task of extracting a single source from a mixture of sources. An example is to separate the signals of specific instruments from a pop music piece. This technique is useful for removing vocal components to create a karaoke sound source, or for creating a score for each instrument in a piece of music. Research on music source separation has been actively conducted to expand the number and types of sources to be separated as well as the accuracy of the separation. Another source separation task is to remove noise from a noisy speaker's speech signal. This paper proposes a method for extracting speech from noisy speech by removing noise.

A neural network propagates data from input to output in a computational manner according to a pre-designed network structure. For many problems, the performance can be improved by designing the structure using prior knowledge. However, it is difficult to improve the efficiency of learning in areas that cannot be compensated by prior knowledge. Self-attention is a method designed to deal with these problems by learning and determining the way the data flows itself, paying attention to the results of its own intermediate calculations, and calculating relevance by paying attention to all positions in the same sequence. It has been applied in fields such as machine translation and image generation [6][8]. Self-attention has also been applied to source separation [3], where each time segment is associated with other time segments that share the same repetitive patterns, and these repetitive patterns are used as additional information for source separation. Self-attention is an attention mechanism that indicates the similarity and importance between the elements of itself, and for each element it calculates the

query Q, key K, and value V using that element (the same values are used for Q, K, and V). When d_k is the dimensionality of the query and key, it is computed as in Eq. (1). The inner product of the query and the key is calculated and divided by the number of dimensions to take into account the context of the whole series, and then the softmax function is applied to prevent the gradient from being lost. Then, by multiplying the calculated weights by the same values as the original input, the output takes into account the context of the training.

$$\mathrm{attention}(Q, K, V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

Self-attention controls which elements are weighted based on the similarity of the query and key. As a result of learning, if it is better to read from an element, the corresponding query and key are updated to be closer, and if it is better not to read from an element, they are updated to be farther apart. In this way, the system automatically decides which element to read.

In [3], self-attention was introduced into a model called Dense-Unet, and it was shown that self-attention leads to improved accuracy in source separation. Dense-Unet was created by synthesizing a structure called Dense Net, which directly connects all layers, and a structure called U-Net, which has skip connections to pass information at each layer between encoders and decoders. Although the Dense-Unet model showed high accuracy compared to previous studies, they showed that the accuracy was further improved by introducing self-attention.

In addition to the above studies, various other methods have been proposed for source separation models. In a study of source separation using a transfer learning approach [5], a model used for speech recognition is trained on a large dataset, and the features are transferred to a source separation model using DenseNet, thereby solving the conventional problem of not being able to maintain long-term dependencies. In this paper, we present an audio query-based separation. In a study of audio query-based separation [2], various types of sound sources are separated by directly compressing the same sound source as the musical instrument to be separated into a latent vector and feeding it to the U-Net model as the target information to be separated.

The purpose of this study is to further improve the accuracy of Open-Unmix, a high-performance and open-source source separation model, by introducing self-attention. Self-Attention is introduced to improve the performance of separation by exploiting long-term internal dependencies when the noise is repeated. Open-Unmix[4] is a three-layer bi-directional LSTM model that takes the spectrogram of the mixed sound source as input and learns to predict the spectrogram of the target sound source for each instrument and vocal of the song. The model learns a mask to remove all sources except the target source, and performs source separation by multiplying the input source by the mask. Among open-source sound source separation software, it shows very high accuracy in the separation results. In addition, we apply the model used for sound source separation of music to the task of sound separation, which is to remove noise from noisy speech information[1], showing that sound source separation research can be applied to various tasks.

[1] https://github.com/seth814/open-unmix-pytorch

2 System Description

In this section, we describe the source separation system proposed in this paper.

2.1 Input Data

In this study, we created a database consisting of speech and noise that can be adapted to the input format of Open-Unmix, a source separation model for music. A mixture of speech and noise was created, and the system was trained to separate them. For noise data, we used the ESC-50[2] dataset. ESC-50 is a dataset of 50 classes and 2,000 files of environmental sounds. ESC-50 is a dataset of 2,000 files of 50 classes of environmental sounds, including animal sounds, rain sounds, human coughs, clock alarms, engine sounds, and other sounds without voices (words). Each file is 5 seconds long and has a sampling rate of 44.1 kHz. It consists of audio data (.wav) and metadata (.csv), and the metadata contains the file name, class (0-49), and class name.

For the audio data, we used the publicly available podcast[3] data. We used the podcast data because the speakers speak clearly and there is little noise. We used the data of 10 broadcasts (95820 seconds).

The data set was pre-processed. First, since each sound source of ESC-50 is 5 seconds long, we converted the podcast audio data into a wav file and divided it into 5-second segments. Then, for each sound source, the data was divided into 80 % for training data, 10 % for validation data, and 10 % for test data. Open-Unmix supports data input in the form of source folders rather than track folders, and the data loader loads random combinations of target and interferograms as input. The model then estimates the mask of the target, and finally outputs the target.

2.2 Proposed Model

The proposed method consists of Open-Unmix and self-attention. These are described by Pytorch[4]. The model structure is as follows (Fig. 1).

The input signal is first converted into a spectrogram by STFT. The input spectrogram is standardized using the mean and standard deviation of each frequency bin over all frames. In addition, batch normalization is applied at several points in the model to stabilize the training. When training with LSTM, the frequency and channel axes of the input information are compressed before training, instead of using the original input spectrogram resolution. This is expected to reduce redundancy and training time. Open-Unmix is composed of three layers of BLSTM. After applying BLSTM, the signal is decoded and returned to its original input dimension. The output is finally multiplied by the mixture of sources as a mask is generated to separate the target sources. In order to perform the separation to multiple sources, the model is trained simultaneously for

[2] https://github.com/karolpiczak/ESC-50
[3] http://podcasts.joerogan.net
[4] https://pytorch.org/
[5] https://github.com/sigsep/open-unmix-pytorch

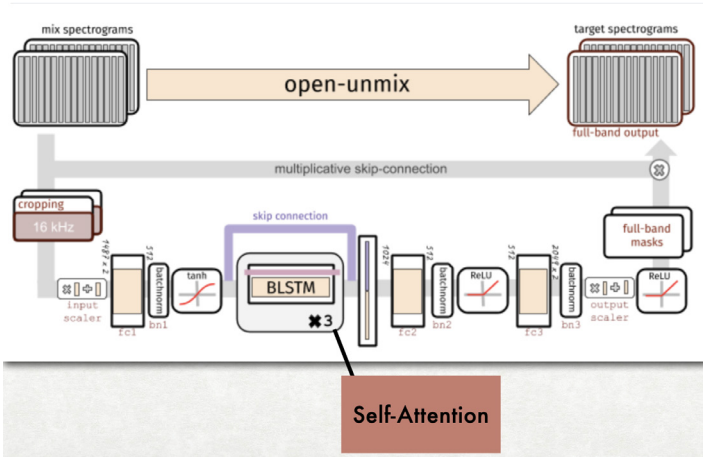

Fig. 1. Diagram of the Open-Unmix[5] with self-attention.

each specific target. In the case of this study, a model for noise extraction and a model for speech extraction are trained simultaneously.

As shown in Fig. 1, we combine self-attention with the output of BLSTM to weight which elements should be focused on during training. The input size to BLSTM is (255,128,512), and the output is returned as a tuple, passing the first element, the hidden layer vector. We concatenate the output of BLSTM and the information held by the skip connection, and the size becomes (255,128,1024). Then, the weighting by self-attention is added.

3 Evaluation Results And Comparisons

3.1 Experiment

We compare the speech separation accuracy of the proposed method with that of Open-Unmix alone. For self-attention, we set the number of channels to be compressed in the convolutional layer to 100 and the output size in the linear layer to 32. The other parameters are batch size 128, window length 512 for STFT, hop count 160 for STFT samples, and data format .wav. The other parameters in Open-Unmix are set by default. SDR, ISR, and SAR were used as evaluation indices [1]. These indices were calculated using the museval package[6]. The units are all expressed in dB, and the larger the value, the higher the accuracy.

3.2 Results

Table 1 shows the results of the objective indices for the test data of the previous and proposed methods. We evaluated the results for both the separated voice and noise.

[6] https://github.com/sigsep/sigsep-mus-eval

The following functions are used to measure performance: Source to Distortion Ratio (SDR), Image to Spatial distortion (ISR), and Sources to Artifacts Ratio (SAR) [1][7].

Table 1. Comparison of previous and proposed methods. Evaluations were calculated for speech and noise respectively.

Metric(dB)	Open-Unmix	Proposed method
SDR (voice)	6.83	7.46
SDR (noise)	8.43	8.59
ISR (voice)	10.24	8.97
ISR (noise)	12.55	12.86
SAR (voice)	6.38	6.53
SAR (noise)	8.92	9.02

All the noises in ESC-50 and the podcast were synthesized, and the evaluation index of speech separation was calculated for each noise and represented using a box-and-whisker diagram.

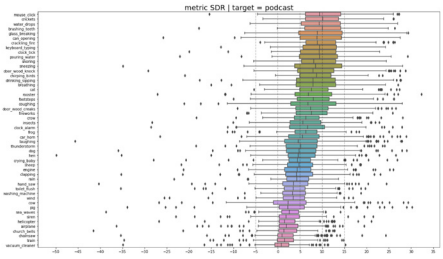

Fig. 2. SDR for all noises (Open-Unmix).

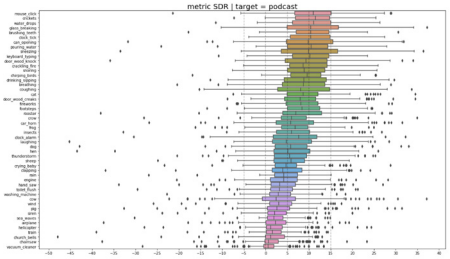

Fig. 3. SDR for all noises (Proposed method).

3.3 Discussion

A t-test was conducted to determine if there was a significant difference in the evaluation of the results. According to Table 2, there was a significant difference in the improvement of accuracy for both voice and noise in SDR, and a significant difference in the improvement of accuracy for voice and noise in SAR. On the other hand, ISR showed a decrease in accuracy in voice.

Since SDR is an overall evaluation value that includes all other evaluation metrics, the improvement in SDR indicates that self-attention is useful for voice separation. On the other hand, the accuracy of the ISR for voice has decreased. It is possible that the weighting of self-attention was not done correctly, or that there was a problem in the model.

Table 2. t-test for evaluation measures.

	SDR(voice)	SDR(noise)	ISR(voice)	ISR(noise)	SAR(voice)	SAR(noise)
t-ratio	-0.62	-1.89	10.35	0.65	1.33	0.1
degree of freedom	199	199	199	199	199	199
significance level	0.05	0.05	0.05	0.05	0.05	0.05
t-distribution	1.65	1.65	1.65	1.65	1.65	1.65
significance of test	P < 0.05	P < 0.05	P > 0.05	P < 0.05	P < 0.05	P < 0.05

The value of SDR changes depending on the type of noise (Figs. 2, 3). This is due to the fact that intermittent noises and noises that are not too loud tend to have higher accuracy than continuous noises during 5 seconds. While none of the previous methods exceed 35 dB, the proposed method exceeds it for three noises.

4 Conclusion

In this paper, we attempted to further improve the accuracy of the Open-Unmix model, which has high performance in open source software, by introducing self-attention. In addition, we demonstrated the versatility of this model for the source separation task by using it not for the source separation task, which separates vocals and accompaniment from a piece of music, but for the speech separation task, which separates each from a mixture of voice and noise. The results of the t-test showed a significant difference.

References

1. Fevotte, C., Gribonval, R., Vincent, E.: BSS EVAL toolbox user guide Revision2.0, Technical report, IRISA (2005)
2. Lee, J. H., Choi, H.-S., and Lee, K.: Audio query-based music source separation, in Proceedings of the International Society for Music Information Retrieval Conference 2019, p. 878–885 (2019)
3. Liu, Y., Thoshkahna, B., Milani, A., and Kristjansson, T.: Voice and ac- companiment separation in music using self-attention convolutional neural network, in arXiv:2003.08954 (2020)
4. Stoter, F.-R., Uhlich, S., Liutkus, A., and Mitsufuji, Y.: Open-Unmix - a reference implementation for music source separation, in Journal of Open Source Software (2019)
5. Takahashi, N., Singh, M. K., Basak, S., Sudarsanam, P., Ganapathy, S., and Mitsufuji, Y.: Improving voice separation by incorporating end-to- end speech recognition, in Proceedings of IEEE ICASSP 2020, pp. 41–45 (2020)
6. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I.: Attention Is All You Need, in Advances in Neural Information Processing Systems, pp. 6000–6010 (2017)
7. Vincent, E., Sawada, H., Bofill, P., Makino, S., and Rosca, J.: First stereo audio source separation evalution campaign: data, algorithms and results, in Proceedings of ICA 2007, pp. 552–559 (2007)
8. Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A.: Self-Attention Generative Adversarial Networks, in International Conference on Learning Representations (2019)