ATTENTION WAVE U-NET FOR AUDIO SOURCE SEPARATION

Yuan Weitao
Adviser

Liu Ding
Lecturer

Li Tingle
Department, Computer Science
Class, CS-1602
Number, 1611640205

Chen Jawei
Department, Computer Science
Class, CS-1602
Number, 1610530404

Liu Cuntian
Department, Computer Science
Class, CS-1601
Number, 1611640105

Abstract: *U-net is an end-to-end network structure that is often used for image segmentation, and it also has applications in audio source separation, but U-net usually uses a direct crop and concat method, which makes the network can not learn enough. In this case, We propose a novel attention gate (AG) model for audio source separation that automatically learns to focus on target structures of varying shapes and sizes. U-net trained with AGs implicitly learn to suppress irrelevant regions in a music while highlighting salient features useful for a specific task. Experiments for audio source separation indicate that our architecture yields a performance comparable to a state of-the-art spectrogram-based U-Net architecture, which gives the same data. Finally, we have made further discussion and prospect to the model.*

Keywords: *Attention Gate, Audio Source Separation, U-net*

1 Introduction

U-net is often used for image segmentation, and it also has applications in speech separation. [1] proposed a novel attention gate(AG) and applied it to medical imaging, which has its unique advantages compared with the previous u-net. The main point is that it combines the underlying and high-level information.

However, this structure (AG) has not been tried in terms of speech, so we migrated this idea to speech and applied it to single-channel speech separation.

We proposed the Attention Wave U-net for end-to-end audio source separation without any pre- or postprocessing, and applied it to singing voice and multi-instrument separation. A long temporal context is processed by repeated downsampling and convolution of feature maps to combine high- and low-level features at different time-scales.

The low-resolution underlying (deep) information after multiple downsampling, can provide contextual semantics for the split target throughout the signal, and this semantic information reflects the characteristics of the relationship between the target and its environment.

After the concatenate operation, the high-resolution (shallow) information from the encoder directly to the high-resolution decoder can provide more detailed features such as gradients.

The reason for adopting this structure is that since it can automatically learn to focus on target structures of different shapes and sizes on the image, thus applying it to speech can also implicitly learn how to suppress irrelevant regions in music and highlight the salient features that are useful for a particular task.

2 Attention WAVE U-NET

We model our system in time-frequency (T-F) domain. Our goal is to separate a single-channel mixture waveform $M \in R^{L_m \times 1}$ into two source waveforms S^1, S^2, with $S^k \in R^{L_s \times 1}$. L_m and L_s as the respective numbers of audio samples. For model variants with extra input context, we have $L_m > L_s$ and make predictions for the centre part of the input.

2.1 The Base Architecture

Wave-U-Net architecture is shown in Figure 1. It uses down-sampling (DS) blocks to compute an increasing number of advanced features on a rougher time scale. These features are combined with local high resolution features, which is calculated previously using up-sampling (US) blocks to generate multi-scale features for prediction. The network has a total L level, and the running time resolution of each continuous level is half that of the previous level. For the K sources to be estimated, the model returns the prediction in the interval (-1,1), one for each source audio sample.

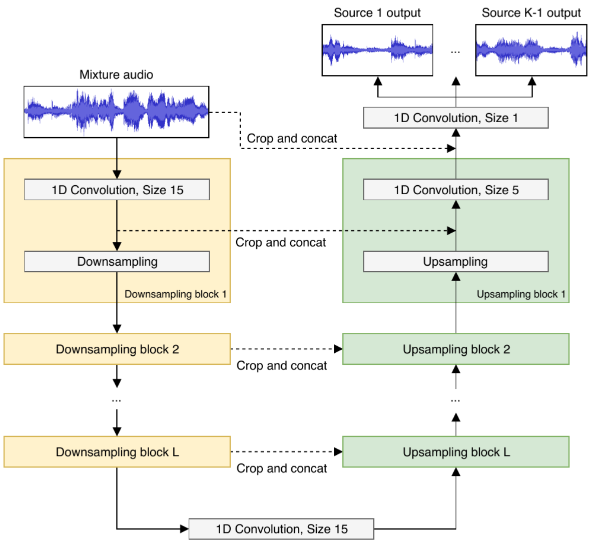

Fig. 1. Base Wave U-Net architecture.

The detailed architecture is shown in Table 1. Conv2D (x, y) denotes a 2D convolution with X filters of size y. It includes zero-padding of the base structure, followed by Leaky ReLU activation (except the last one using tanh). Round the discard function for each other time step to halve the time resolution. Upsample performs twice as much as the sampling in the temporal direction, and we use linear interpolation. Concat (x) connects current advanced functions to more local functions X. In extensions of the base structure extensions, Conv2D does not involve zero-padding, and X first performs central tailoring, so it has the same number of time steps as the current layer.

2.2 Architectural Improvements

We can find that the structure shown in Figure 1 is passed directly from the encoder to the high-resolution (shallow) information on the same height decoder. In our structure, we introduce the attention gate (as shown in the Figure 2), which is a structure similar to a deep residual network.

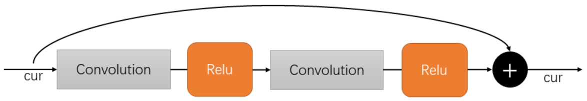

Fig. 2. Res-net structure diagram.

Attention gate is formulated as follows (as shown in the Figure 3):

$$\begin{cases} \alpha = relu\left(conv\left(cur\right) + conv\left(pre\right)\right) \\ \beta = sigmod\left(conv\left(\alpha\right)\right) \\ cur = \beta \cdot cur \end{cases}$$

Why should we adopt such a structure? The preliminary introduction has already explained its application in image segmentation. We know that upsampling can complement some signal information, but the information supplement is definitely not complete, so it also needs to be sampled with the left side. The higher signals are connected, copied directly and then cropped to the same size as the upsampled signal. In order to improve the accuracy, further make a compromise between the high sampling rate and the more abstract features, and draw on the idea of image segmentation to attention gate, the output of AG is the element-wise multiplication of input feature-maps and upsampling layer. Because as the number of convolutions increases, the extracted features are more efficient and more abstract. The upsampled signal is a signal that undergoes multiple convolutions. It is definitely a more efficient and abstract signal, and then connected to the less abstract but higher-resolution feature on the left to get a more complete signal.

The advantage of our structure is that it preserves both the U-net direct connection structure and the Res-net addition structure, and we can understand this operation as image data argumentation processing, although it is applied audio aspect

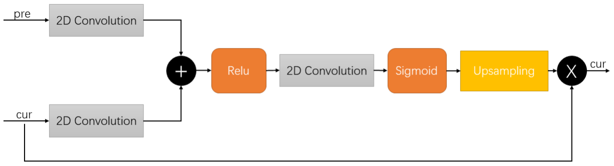

Fig. 3. Schematic of the proposed additive attention gate (AG).

3 Experiments

We evaluated the performance of our model on two tasks: using vocals and accompaniment as a classification to audio source separation.

3.1 Datasets

The 75 tracks from the training zone of the MUSDB [3] multi-track database are randomly assigned to our training set, and the remaining 25 tracks form the validation set for early stop. Final performance was evaluated on a MUSDB test partition containing 50 songs. For the separation of voices, we also added the entire CCMixter database [4] to the training set.

We multiply the source signal by the factor uniformly selected from the interval [0.7, 1.0] as the data augmentation for the two tasks, and set the input mix to the sum of the source signals. No further data preprocessing is performed, only conversion to mono (except stereo model) and downsampling to 22050 Hz.

3.2 Training Procedure

During training, the audio excerpts are randomly sampled and the inputs are filled accordingly.Enter the context model. As loss, we use mean squared error (MSE) over all the batch source output samples. We use the ADAM optimizer with learning rate 0.0001, the decay rates β1 = 0.9 and β2 = 0.999 and the batch size is 16. We define 2000 iterations as one epoch, and if the validation set is not improved after 20 epochs which is measured by MSE loss,we perform the early stop. Then the last model goes one fine-tuning stage, which is twice as much as the batch size before, the learning rate is reduced to 0.00001, and then if the validation loss was not improved in the next 20 epochs, the model with the best validation loss will be selected.

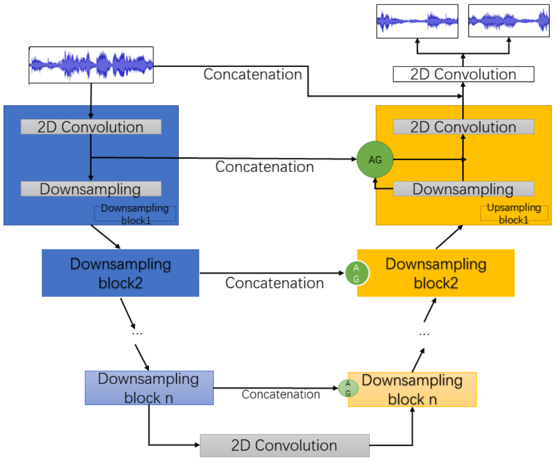

Fig. 4. Our proposed Attention Wave U-Net with n layers.

3.3 Model Settings and Variants

For our baseline model, we use $L_m = L_s = 16384$ input and output samples, $L = 12$ layers, $F_c = 24$ additional filters, Each layer and filter size is $f_d = 15$ and $f_u = 5$.

We trained the U-Net architecture based on the spectrogram [5](original). Because each sample requires a lot of memory, the initial batch size is set to 4. To train original model, we operate the backpropagation error by inverse STFT, which is used to construct the source audio signal based on the estimated spectrogram amplitude and the mixed phase. Although we strive to make a holistic model comparison, please note that some training settings (such as the learning rate used in [5] may be different (and partially unknown) from ours, and can provide better performance than U7, and not shown here, even the same dataset.

4 Results

4.1 Evaluation metrics

The signal-to-distortion (SDR) metric is commonly used to evaluate source separation performance [6]. but the mean and standard deviation are not sufficient to adequately summarise the vocal[7], As a workaround, [7] use the median absolute deviation (MAD) as a rank-based equivalent to the standard deviation(SD). It is defined as the median of the absolute deviations from the overall median and is easily interpretable, sincea value of x means that 50% of values have an absolute difference from the median that is lower than x

4.2 Model comparison

origianl: WAVE U-NET
ours: Attention WAVE U-NET

		original	ours
Voc.	Med.	2.76	2.86
	MAD	2.46	2.43
	Mean	-0.66	-1.44
	SD	12.38	13.99
Acc.	Med.	6.76	6.78
	MAD	2.00	1.97
	Mean	6.90	6.92
	SD	3.67	3.53

Table 1. Test set performance metrics (SDR statistics, in dB).

5 Discussion and Conclusion

In our preliminary experiments, the performance of our model seems to be worse than the original model, but this does not mean that our network architecture is worse than the most advanced architecture, because the training set was unexpectedly stopped during the experiment. Improvement, so the evaluation model we got is not the best of this experiment. Due to the time, the fine-tuning measures that should be taken after training are also forced to be omitted.

As highlighted in previous papers, the lack of proper time input context in recent separation and enhancement models can compromise performance and create artifacts. In this experiment, we propose a simple change by focusing on the convolution fill as a solution.

For the future work, of course, first of all, continue to adjust the parameters, and strive to get a better result. We are modeling in the frequency domain this time, and the subsequent application will be applied to the time domain. In addition, if allowed, we can study The better loss function of the original audio prediction, such as those provided by the generated confrontation network [8, 9], because the MSE may not reflect the perceived quality loss well.

References

[1] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, Daniel Rueckert. Attention U-Net: Learning Where to Look for the Pancreas. arXiv preprint arXiv:1804.03999, 2018

[2] Daniel Stoller, Sebastian Ewert, Simon Dixon. Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation.arXiv preprint arXiv:1806.03185, 2018

[3] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stter, Stylianos Ioannis Mimilakis, and Rachel Bittner. The MUSDB18 corpus for music separation, 2017.

[4] Antoine Liutkus, Derry Fitzgerald, and Zafar Rafii. Scalable audio separation with light kernel additive modelling. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 76–80. IEEE, 2015.

[5] Andreas Jansson, Eric J. Humphrey, Nicola Montecchio, Rachel Bittner, Aparna Kumar, and Tillman Weyde. Singing voice separation with deep U-Net convolutional networks. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 323–332, 2017.

[6] E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. IEEE Transactions on Audio, Speech, and Language Processing, 14(4):1462–1469, 2006.

[7] Vincent Dumoulin and Francesco Visin. A guide to convolution arithmetic for deep learning. arXiv preprint arXiv:1603.07285, 2016.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in Neural Information Processing Systems, pages 2672–2680, 2014.

[9] Daniel Stoller, Sebastian Ewert, and Simon Dixon. Adversarial semi-supervised audio source separation applied to singing voice extraction. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 2391–2395, Calgary, Canada, 2018. IEEE.