Sound event detection and localization based on CNN and LSTM

Technical Report

Zhao Lu

University of Electronic Science and Technology
of China.
School of Information and Communication
Chengdu, China
Zlu40@qq.com

ABSTRACT

The Detection and Classification of Acoustic Scenes and Events (DCASE) 2019 challenge is a topic seminar for speech feature classification. Task 3 is the location and detection of sound events. In this field, the learning method based on deep neural network is becoming more and more popular. On the basis of CNN, the spectrum and cross-correlation information of multi-channel microphone array are learned based on CNN and LSTM, and the detection of sound events and the estimation of arrival direction are obtained. Compared with the baseline method, this method improves the estimation accuracy of DOA and the recognition ability of SED by using DCASE2019 dataset and PyTorch deep learning tool. The combination of CNN and LSTM works very well on this kind of feature classification problem with time series.

***Index Terms*—**Convolutional neural network, Sound event detection, direction of arrival, Long Short-Term Memory,

1. INTRODUCTION

The localization and detection of sound events includes two sub-branches of sound event detection and sound arrival direction positioning. A SELD task system based on microphone array signals is given in Task3 of DCASE2019[1]. According to this system, we can verify our own implementation of SELD detection. And compare with standard data to assess the overall performance of the system. The basic goal of sound event detection and positioning needs to be achieved is to determine the type and start end time of each event in the environment, and to determine the direction of arrival in space for each sound event, that is, to give an estimate of the azimuth and elevation angles.

The application of SELD system can automatically realize the perception and detection of sound field environment in space, and make the intelligent machine have the same ability of sensing and detecting sound in space as the human ear. It has been widely used in active noise reduction, sound and speech enhancement, speech recognition, sound information visualization, intelligent interaction, robot perception and other applications.

2. FEATURE EXTRACTION

The dataset applied to this task is the dataset of TAU Spatial Sound Events 2019-Microphone Array[2]. The dataset provides a four-channel directional microphone recording configured by a tetrahedral array, which can collect sound information from the entire space. The dataset contains a development and evaluation set. There are 400 clips of sound. Each sound file includes four channels of recording waveforms, and the corresponding CSV file contains the recording of sound events contained in each recording clip, the start and end time of each sound event, the azimuth and pitch angle of the direction of arrival of the sound source. The angle is divided into 10 degrees.

The main purpose of Mel feature spectrum extraction is to convert the actual spectrum into the frequency which is easier to be perceived by the human ear[3]. The conversion formula (1) is as follows

$$Mel(f) = 2595\log(1+f/700) \tag{1}$$

The basic processing process is dividing the audio signal into frames, and then making the STFT to each group of signals to get the spectrum information, and then getting through the Mel filter, then taking the logarithm of the results. the Log-Mel characteristic spectrum of the current channel signal can be obtained. In the process of subsequent speech processing and training, the feature information is used as the input feature.

From the calculation process of Log-Mel spectrum, we can know that each calculation is to calculate the information of each channel separately. For microphone arrays, the time difference between them is not utilized. For the estimation of DOA, the most important task is to accurately estimate the time difference between the sound source and each sensor. In order to make better use of the relationship between the information received by the microphone array, it is necessary to calculate the cross-correlation function between each group of data. The spectral information of GCC, that is, the delay information between different microphones, is obtained. For four-channel microphones, we can get six sets of delay information. Using PHAT as a weighted function, the simple form of calculating GCC-PHAT can be expressed as follows (2)

$$\psi_{i,j}^{PHAT}(t,\tau) = IFFT\{\frac{X_i(f,t)[X_j(f,t)]^*}{|X_i(f,t)[X_j(f,t)]^*|}\} \tag{2}$$

The GCC-PHAT spectrum can be obtained by calculating the cross-correlation power spectrum of the two groups of signals, multiplied by the weighted function and inverse Fourier transform. [4]The vector contains the time difference information of different microphone signals. GCC-PHAT is also a common method to solve DOA. Here, it is mainly used as input feature vector for training.

3. MODEL TRINING

In the field of sound event detection and DOA estimation, the deep learning method based on CNN has been widely used. Convolution neural network has a strong performance in the field of image classification and recognition. For sound events, the input spectrum information is also a kind of two-dimensional image information for the computer. The learning and classification of depth features for this kind of information is the same as the CNN of ordinary images. Therefore, we can use the specific atlas of different sound events to learn, in the training process, take the accurate sound events as the learning goal, optimize our network parameters. To achieve the purpose of event classification. In addition to CNN, cyclic neural network RNN is also a common deep network model. Unlike CNN, RNN has strong performance in processing sequence data. Therefore, RNN is also widely used in the field of natural language processing. LSTM (Long Short-Term Memory) is a long-term and short-term memory network, which is a time-cyclic neural network, which is suitable for processing and predicting important events with relatively long intervals and delays in time series. The main purpose of this network is to deal with the problem of gradient disappearance and gradient explosion in the training process of RNN network. To put it simply, it can perform better in longer sequences than a normal RNN, LSTM. In SED and DOA tasks, the input spectrum information is not a simple two-dimensional spectrum, but a spectral information with temporal relationship. The information of each moment and its sequence are related to each other. Using the time-related network, the order of input information can also be studied and trained. Compared with simple CNN network, the network combining CNN and RNN has stronger ability to process time information. Therefore, we use CNN+LSTM network as the training network model.

Our network mainly includes input layer, CNN layer, dropout layer, LSTM layer and fully connected layer. Finally, an estimate of the two directional angles of the SED and DOA is obtained. PyTorch was used as a framework for deep learning. As the most popular deep learning framework, PyTorch not only has a simple and clear structure, but also provides users with a lot of detail adjustment space. And it has strong scalability and a lot of learning resources. This tool can efficiently implement various deep learning tasks.

The overall network structure consists of three major parts. The first part is the 8-layer CNN layer, which mainly performs deep feature extraction, Dropout layer, LSTM layer and fully connected layer.

The CNN layer has a kernel size of 3x3 and the output layer number is 64, 128, 256, 512. The activation function is the Relu function. The LSTM uses a bi-directional network with an input of 512 and an implicit layer of 256, a total of two layers of Num size. Finally, there is a 512-dimensional fully connected network with an output dimension of 11. The SED is estimated using the Sigmoid function, and the DOA directly uses the linear activation function. The sketch of the network is shown in Figure 1.

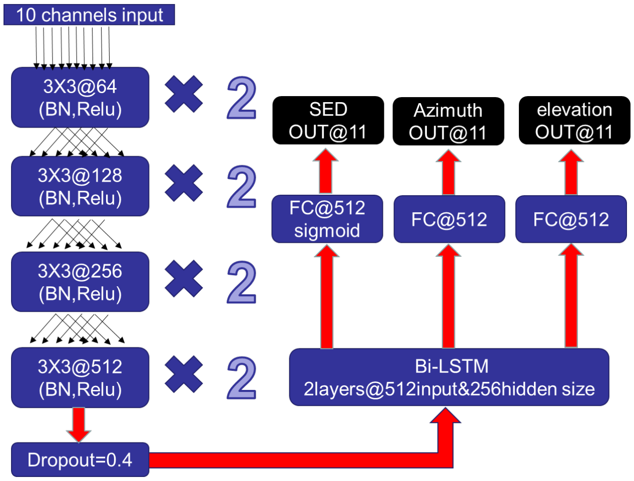

Figure 1: Basic structure of the CNNLSTM network

4. TEST AND ANALYSIS

The sampling frequency of our feature extraction is fs = 32000Hz, the number of Mel filters is 1024, the number of FFT points is 1024, the number of Mel filters is Mel bins = 64, and each second is divided into 64 units. The STFT result is calculated and the Log-Mel spectrum is obtained. Then, the GCC result is calculated for the signals of the four channels, and the correlation coefficient is calculated for the four channels, and six sets of results can be obtained. We can get 10 sets of input feature vectors. Used as a feature data set for training. For a given dev data, 400 sets of data are divided into four sets, 1, 2 for the training set, 3 for the test set, and 4 for the validation set. Each collection contains 100 sets of data. Train with the Adam optimizer. A total of 50 groups of epochs were trained to observe the change in error rate. In fact, after 15 groups of epochs, the trainer converges to a more stable situation. Run the test system and compare it with the standard data set to get the results of the evaluation. They are SED error, F score, DOA error, DOA recall and SELD point. From these results, we can evaluate the performance of our classifier.

DCASE gives a visual tool that allows us to visualize the performance of the classifier. On the left is the result of the standard data set, and on the right is the result of the trained classifier. It can be seen that the classifier can estimate the classification and duration of the sound events, and the direction of arrival of each sound

5. RESULTS AND ANALYSIS

Compared with the current method with SELD net's Baseline method[5], 9-layer CNN [6]and CNN+GRU [7]using Log-Mel spectra as input features, the test data is the official data of the TUT dataset. This method has a significant improvement in the detection accuracy of DOA, and also has good results in the detection accuracy of SED. Table 1 shows the results of different methods.

Method	feature	SED error rate	F1-score	DOA error	DOA frame recall	SELD score
SELDnet	STFT	0.366	0.815	27.140	0.829	0.218
CNN_9Layer	Log-Mel	0.269	0.844	18.784	0.824	0.176
CNN+GRU	Log-Mel	0.193	0.881	16.492	0.850	0.138
CNN+LSTM	**Log-Mel+GCC**	**0.159**	**0.904**	**7.175**	**0.852**	**0.111**

Table 2: The comparison of different network

Using DCASE-SELD's visualization tool, you can visually display the comparison of test results with accurate data, including sound events and start and end times. The ordinate is divided into 11 parts, which represent the specific events in the data in 11 and the corresponding two events. Angle information. It can be seen that this method has a good detection effect for the information of SED and DOA. The Figure 2 shows the visualization of the result.

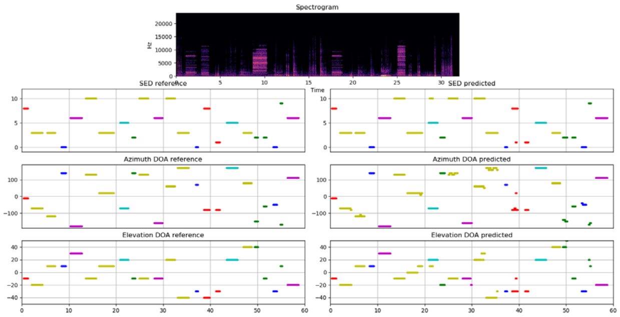

Figure 2: The Visualization of the result

The main deviations occur in complex situations, polyphonic events with multiple source changes. At this point Azimuth's estimate will be offset. Elevation's estimate is slightly better, but the direction of certain events will be inaccurate. Some SED estimates may also have estimated or more estimated events, and further improvements are needed for environments with multiple source aliasing. Figure 3 shows the result in polyphonic sound events.

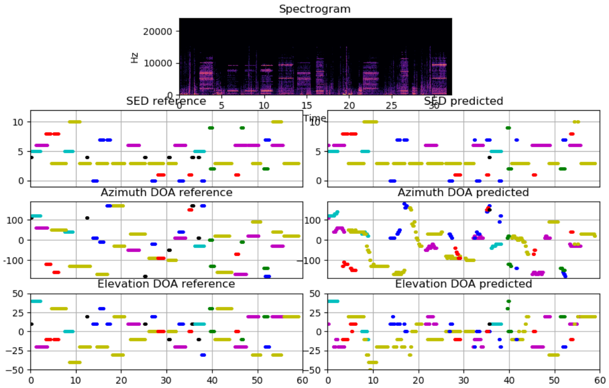

Figure3 : The result in polyphonic sound events

6. REFERENCES

[1] http://dcase.community/workshop2019/.

[2] https://github.com/sharathadavanne/seld-dcase2019

[3] S.B. Davis, and P. Mermelstein, “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences,” in IEEE Transactions on Acoustics, Speech, and Signal Processing, 28(4), pp. 357–366,1980.

[4] C. Knapp and G. Carter, “The generalized correlation method or estimation of time delay,” IEEE Transactions on Acoustics, speech, and Signal Processing, vol. 24, no. 4, pp. 320–327,1976.

[5] Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. IEEE Journal of Selected Topics in Signal Processing, ():1–1, 2018.

[6] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yong Xu, Wenwu Wang, Mark D. Plumbley. Cross-task learning for audio tagging, sound event detection and spatial localization: DCASE 2019 baseline systems. arXiv preprint arXiv:1904.03476 (2019).

[7] Yin Cao, Qiuqiang Kong, Turab Iqbal, Fengyan An, Wenwu Wang, Mark D. Plumbley. Polyphonic Sound Event Detection and Localization Using Two-Stage Strategy. arXiv preprint arXiv: 1905.00268v2 (2019).