Environmental sound classification based on feature fusion

Huimin Zhao[1,*], *Xianglin* Huang[2] ,*Wei* Liu[3], and *Lifang* Yang[4]

[1]Communication University of China, Department of Computer Science, Beijing, China
[2]Communication University of China, Department of Computer Science, Beijing, China
[3]Communication University of China, Department of Computer Science, Beijing, China
[4]Communication University of China, Department of Computer Science, Beijing, China

Abstract. With deep great breakthroughs of deep learning in the field of computer vision, the field of audio recognition has gradually introduced deep learning methods and achieved excellent results. These results are mainly for speech and music recognition research, and there is very little research on environmental sound classification. In recent years, people have begun to expand the research object of deep learning to the environmental sound, and achieved certain results. In this paper, we use ESC-50 as our test set, based on the SoundNet network and EnvNet network to propose a feature fusion method[1]. After the features extracted by SoundNet and EnvNet were merged, they were classified using fusion features. Experimental results show that this method has better classification accuracy for the recognition of environmental sounds than using either of the two networks separately for classification.

1 Introduction

Environmental sound is a very rich audio that occurs in our daily lives. It is a special kind of audio that differs from voice and music. For example, the sound of animals in nature, the sound of opening doors in our lives, and the sound of planes taking off at airports. and many more. However, due to the complexity of environmental sounds, limited data volume and other reasons, research on environmental sound recognition has lagged far behind the study of speech and music[2]. Environmental sound recognition is a very effective method for sensing the surrounding environment and can be widely used in many fields such as driverless, mobile robots, wearable devices, and smart homes[3]. In recent years, deep learning has provided tremendous potential for the development of artificial intelligence. Computers have taken on more and more complex tasks.

Traditional methods for ambient sound recognition are based primarily on hand-designed audio features such as MFCC, zero-crossing rate, short-term energy, and so on. These methods only consider one or more aspects, and do not fully consider the richness and complexity of environmental sounds[1]. Their performance is still far from the ideal level. In order to further enhance the recognition performance of environmental sounds, deep learning techniques have been introduced. Deep neural network in deep learning as a high-performance, multi-level neural network has been widely proved to have certain advantages in extracting data features and establishing recognition models. It can extract features that humans cannot design from raw data. . In the field of image recognition, the neural network can directly extract features from the original image; in the field of audio recognition, the neural network can learn sound directly from the original waveform and extract features other than human designed features such as MFCC[4].

Based on the SoundNet network proposed by Yusuf Aytar et al. and the end-to-end EnvNet network proposed by Yuji Tokozume et al., we identified the environmental sound. The pre-processed data was sent to the SoundNet network and the EnvNet network after certain preprocessing. The audio features extracted from the SoundNet network were combined with the audio features extracted from the EnvNet, and the fused audio features were used as the training and test data of the classifier. We use the ESC-50 data set to evaluate performance. The experimental results show that the accuracy of the fusion of SoundNet and EnvNet has improved by 3.2%. This result shows that the fused network can help improve the classification accuracy

2 Dataset

The ESC-50 dataset is a balanced 50-item, 2000-length, 5-second recording of the environment selected from 5 categories (animal, natural soundscape, human nonverbal sound, indoor/family sound, and external/urban noise). Collection. There are 40 samples for each category, and the sampling frequency of the data is 44.1 kHz. The accuracy of this data set for untrained people was 81.3%[1].

In order to train the model more effectively, we made a certain adjustment to the sampling frequency of the original data of 44.1 kHz, and the sampling frequency of the adjusted data was 16 kHz[3].

[*] Corresponding author: zhaohuimin_f@163.com

2.1 Dataset preprocessing

Environmental sounds and other audio such as music, speech have many different places. The ambient sound can generally be of three types: single sounds such as a mouse-click, detailed sounds such as Dog's barking, and steady contimuous sounds such as the sound of washing machine or engine. If the audio duration is too short, it will be difficult Distinguish between single sounds and repeated sounds. If the audio duration is too long, the data will contain a lot of redundancy due to the large amount of silence and repetition[3]. According to the experimental results mentioned in Yuji Tokozume's paper， as shown in the following Fig. 1.[3]. the accuracy of the classification of the ambient sound is 1.5 seconds[3], and the ESC-50 dataset is composed of 2000 sounds with a duration of 5 seconds, so we need to use ESC- The environmental sounds in the dataset are pruned. We use the sliding window method to organize the data. Each sliding step of the sliding window is 0.2s[3].

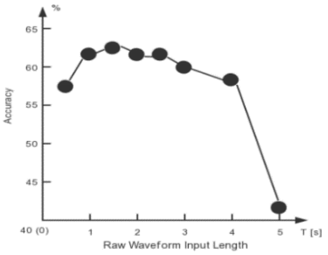

Fig. 1. Relationship of raw waveform input length on accuracy.

The windowing operation of the audio raw waveform is equivalent to multiplying the audio raw waveform with the window function, and observing the waveform through the window function, as shown in the following Fig. 2.

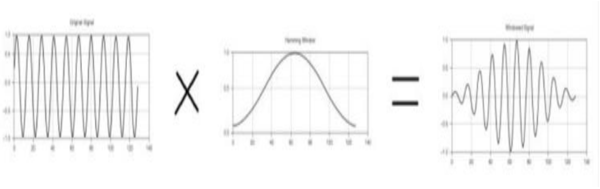

Fig. 2. Waveform through the function schematic

2.2 Window function

The commonly used window functions are Boxcar windows, Fejer windows, Hanning windows, etc[5]. Each type of window function has its own advantages and disadvantages. We have used four kinds of window functions to compare the performance. These are Boxcar windows, Fejer windows, Hanning windows, and Hamming window.

(1) Rectangle window: Rectangle window belongs to the time variable zero-power window, its time function is:

$$\mathrm{w(t)} = \begin{cases} \frac{1}{T}, 0 \le |t| \le T; \\ 0, \quad |t| > T \end{cases}$$

Its time and frequency waveform as shown in the following Fig. 3.

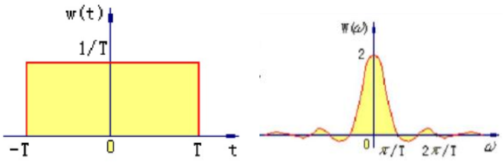

Fig. 3. Rectangular window waveform and spectrum.

Rectangle window is used most, its advantage is that the main lobe is more concentrated, the disadvantage is that the side lobes are high, and there are negative side lobes, leading to high-frequency interference and energy leakage in the transformation, and even a negative spectrum phenomenon[6].

(2) Triangular window: Triangular window is a square form of the power window, its time function is:

$$\mathrm{w(t)} = \begin{cases} \frac{1}{T}(1 - \frac{|t|}{T}), 0 \le |t| \le T; \\ 0, \quad |t| > T \end{cases}$$

Its time and frequency waveform as shown in the following Fig. 4.

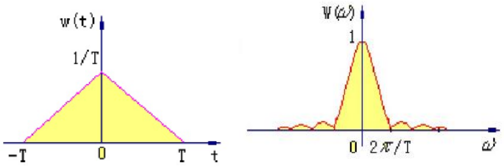

Fig. 4. Triangular window waveform and spectrum.

(3) Hanning window: Hanning window is also called as a cosine window. Its time function is:

$$\mathrm{w(t)} = \begin{cases} \frac{1}{T}\left(\frac{1}{2} + \frac{1}{2}\cos\frac{\pi t}{T}\right), \quad 0 \le |t| \le T; \\ 0, \quad |t| > T \end{cases}$$

Its time and frequency waveform as shown in the following Fig. 5.

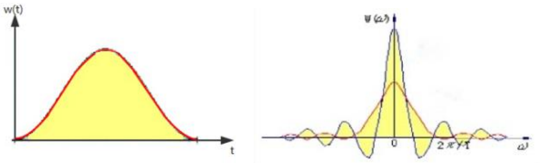

Fig. 5. Hanning window waveform and spectrum.

Hanning window can be regarded as the sum of the spectra of the three rectangular windows. The Hanning window and the rectangular window are compared. From the viewpoint of energy leakage, the Hanning window is better than the rectangular window, but the main lobe of the Hanning window is widened, which is equivalent to the analysis bandwidth[7]. Widening, the frequency resolution decreases.

(4) Hamming window: hamming window is also a raised cosine window, which differs only in the weighting factor from the hanning window[8]. The weighting factor of the Hamming window can make the side lobe smaller.

2.3 Dataset Reconstruction

An important feature of the neural network model performance is that the larger the amount of data involved, the better. The ESC-50 contains a total of 2000 5s audio. We use a window with a length of 1.5s and a sliding step of 0.2s for the original waveform with a duration of 5s. After the function is truncated, the amount of audio data for each class will change from 40 to 40*[[(5-1.5)/0.2]+1]=760; The audio data volume of all 50 classes will change from 2000 to 2000*[[(5-1.5)/0.2]+1]=38000 audio data This operation can appropriately expand the amount of data and improve the accuracy of the classification.

Each window function has its own advantages and disadvantages. When using it, multiple sets of experiments are needed to find the best window function for the data. The experimental results in Table 1 show that using the Hanmming window for dataset reconstruction is much better.

Table 1. The relationship between different window functions and classification accuracy

Network Window function	SoundNet+EnvNet
boxcar	75.2%±0.6%
Bartlett	74.5%±0.5%
hanning	75.9%±0.3%
hamming	77.4%±0.5%

3 Feature extraction and fusion

Different neural network structures have advantages and disadvantages. Different network structures using the same data set will have different results. We should use the most suitable network structure for feature extraction.

3.1 SoundNet

SoundNet is a student-Teacher training procedure proposed by Yusuf Aytar et al., which transfers discriminative visual knowledge from well established visual recognition models into the sound modality using unlabeled video as a bridge. SoundNet utilizes a large number of unlabelled collections from natural environments. Sound data to learn the natural sound representation, and using the synchronization characteristics between the visual and sound of the video, using the migration learning method to learn acoustic representations with 2 million unlabeled videos. SoundNet train deep sound networks by transferring knowledge from established vision networks and large amounts of unlabeled video[2].

The SoundNet network structure is shown in the Fig. 6[2].

For SoundNet networks, it uses discriminative visual knowledge from well established visual recognition models into the sound modality using unlabeled video as a bridge. The advantage of untagged videos is that they can also be made large-scale with limited economic capabilities. Information data. SoundNet uses a large amount of untagged video that contains rich information to migrate discriminating visual knowledge from better visual recognition models to sound forms, greatly improving the accuracy of using only audio knowledge for environmental sound classification. Compared to EnvNet, SoundNet does not take into account the best duration issues for ambient sound classification and audio-specific time domain issues, so SoundNet lacks the characteristics of the time domain.

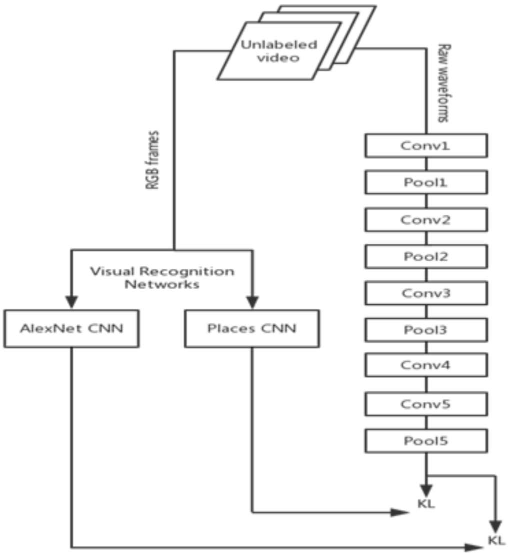

Fig. 6. SoundNet Architecture.

3.2 EnvNet

EnvNet is an end-to-end neural network for identifying ambient sounds proposed by Yuji Tokozume et al.. EnvNet first maps one-dimensional audio data to two-dimensional audio data through a convolutional layer, and then treats the mapped two-dimensional audio data as The classification is done like a picture by a convolutional [3].

The EnvNet network structure is shown in the Fig. 7[3].

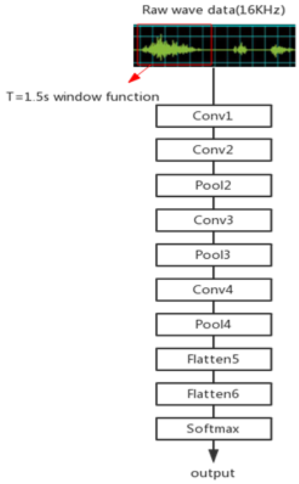

Fig. 7. EnvNet Architecture[3].

EnvNet is an end-to-end environmental sound classification system that transforms one-dimensional audio classification into two-dimensional image classification problems. This operation provides us with more ways to solve problems. EnvNet not only explores the optimal duration for ambient sound classification, but also considers the very important time domain issues in the audio classification field. It also proposes the use of convolutional layers with a very small filter size for ambient sound classification. Compared to SoundNet, EnvNet does not use the rich visual knowledge of video models, so there is a slight lack of model performance.

3.3 Feature extraction and fusion

Considering the advantages and disadvantages of the two networks, we propose a method of feature fusion. We use the SoundNet network model and the EnvNet network model to extract features of the original waveform of the environmental sound, and then fuse the features extracted from the two networks, and then use the fused features to train the classifier. The Soundnet network has two types of network structures, namely, a 5-layer network structure and an 8-layer network structure. The 8-layer network structure adds three-layer convolutional layers to the 5-layer network structure. According to the experimental results mentioned in the Yusuf Aytar paper, we can understand that the 5-layer SoundNet network is more suitable for ESC-50 classification, and AlexNet is a better teacher net work for ESC-50, so we chose AlexNet as the 5th layer of the teacher net[2]. SoundNet network and EnvNet network extract the sound characteristics of the environment. Connect the audio features extracted from the two networks and classify them using SVM. The model was evaluated with a 5-fold crossvalidation scheme.

Multi-network feature fusion Architecture is shown in the Fig. 8.

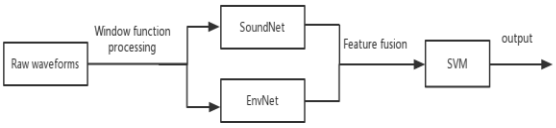

Fig. 8. Multi-network feature fusion Architecture.

4 Conclusions

From the experimental results as shown in table2, it can be concluded that each neural network has its focus on its extracted features. The accuracy of the audio features after the fusion of SoundNet and EnvNet two neural networks is the highest, 3.2% more than the SoundNet network. We analyze that if we take into account the static log-mel feature and delta log-mel feature, we will achieve better performance.

Table 2. The relationship between network structure for feature extraction and classification accuracy

Network / Dataset	ESC-50
SoundNet	74.2%
EnvNet	64%±2.4%
SoundNet+EnvNet	77.4%±0.5%

References

1. Karol J Piczak. 2015, ESC: Dataset for environmental sound classification. in Proc. ACM International Conference on Multimedia, 2015, 9(2):1015-1018.

2. Yusuf Aytar, Carl Vondrick, Antonio Torralba. 2016 "SoundNet:Learning Sound Representations from Unlabeled Video," in Proc. NIPS 2016.
3. Yuji Tokozume, T Harada. 2017 "Learning environmental sounds with end-to-end convolutional neural network," in Proc. ICASSP, 2017, 2721-2725.
4. Karol J Piczak. 2015, Environmental sound classification with convolutionalneural networks. In Proc. MLSP, 2015, 1-6.
5. Qiuying Shi. 2016 . Deep learning-based and transfer learning-based enviorment sound recognition. in Proc. 2016
6. De Liang. 2015. Deep Neural Networks for Chinese Speech Recognition. In Proc. 2015.
7. Lingli Ling. 2011. Environmental Sound Classification Based on HMM and SVM. In Proc. Computer Era, 59-61, 2011.
8. Qingqing Yu, Ying Li, Yong Li. 2011. Natural sounds recognition using GMM distribution. In Proc, Computer Engineering and Applications, 2011,47(25): 152-155.
9. Burak Uzkent, Hakan Cevikalp, Buket D. Barkana. 2012. Non-speech environmental sound classification using SVMs with a new set of features. In Proc, IJIC.