Anomalous Sound Detection Based on Convolutional Neural Network and Mixed Features

Jie Zhao
School of Physics, University of Electronic Science and Technology of China, Chengdu 610054, China
Email: Jacob-zh@outlook.com

Abstract. In this paper, convolutional neural network (CNN) is used to detect abnormal sound. Aiming at the problem that the audio time-frequency graph contains complex feature information and the single traditional audio feature contains insufficient information, a new feature graph is proposed, which can be used as the input of the CNN model to get more accurate detection results. Firstly, the two-dimensional time-frequency graph is obtained by the short-time Fourier transform of the abnormal sound signal. Secondly, this paper extracts audio features such as MFCC and short-term energy from the time-frequency map and the original audio. Finally, this paper uses matrix transformation to combine the features and takes the processed feature graph as the input of the CNN model. And the CNN model was trained to judge the category of abnormal sound. The results of measured data show that the combination of CNN and mixed features can accurately judge and classify abnormal sounds.

Keywords. Audio classification; CNN model; audio feature; sound dataset.

1. Introduction

Nowadays, various kinds of sounds are prevalent in all aspects of people’s life, and some abnormal sounds have the function of warning people of danger. Therefore, the detection of abnormal sounds is becoming an important content in the field of sound research. The detection of abnormal sounds is actually to classify and identify abnormal sounds in life, such as the sound of broken glass, gunshots, screams, etc.

The current researches on abnormal sound detection mainly focus on two aspects, extracting the features of different kinds of abnormal sounds and researching new recognition algorithms. There are many classification features of sound, which can be summarized into three categories: time-domain features, frequency-domain features, and homomorphic features. Although the algorithm for extracting time-domain features is simple, their ability to represent differences is poor. The frequency domain features are more in line with the auditory characteristics of human ears, but the ability to express difference is not so good for the productive combined sound signals. The homomorphic feature can just solve the shortcomings of the frequency feature. At present, the commonly used traditional sound features mainly include zero crossing rate, short term energy, short term autocorrelation coefficient, speech spectrum, MFCC and so on. At the same time, the traditional recognition algorithms include Gaussian Mixture Model (GMM), Hidden Markov Model (HMM) [1], and Dynamic Time Warping (DTW). At present, many articles about abnormal sound detection are implemented based on the above conventional methods. For example, a system for sound event classification based on MFCC and HMM [2, 3]. But because the sound features are single, its classification effect is not very good for abnormal sounds with close features. Moreover, simply increasing the complexity of the transition probability

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd
matrix in this system cannot improve the fitting ability of the model. Valenzise et al. proposed an audio-based monitoring system applied to public squares, which uses two parallel GMM to identify gunshots and screams from the noise environment [4]. The problem with this system is that increasing the order of GMM does not improve the accuracy of the model in recognizing abnormal sounds. With the development of machine learning and the rise of deep learning, the application of deep neural network (DNN) in voice recognition has been further studied. Convolutional neural network (CNN) is a practical model developed in recent years, especially in the field of pattern classification [5, 6]. Due to the huge advantages of CNN in distinguishing complex shapes of images, In 2015, Zhang used spectrum-based image features in the sound event recognition process of convolutional neural networks, which also showed good performance under noise interference [7]. However, when there is noise, the spectrum contains too much miscellaneous information, so the training process is tortuous.

To solve the above problems, this paper applies convolution network (CNN) to abnormal sound detection. This paper proposes a new feature image and uses it as the input of CNN to get more accurate detection results. Firstly, the two-dimensional time-frequency map is obtained by using the short-time Fourier transform of the sound signal. Secondly, MFCC and other audio features are extracted from time-frequency map and original audio. Finally, the features are combined by matrix transformation, and the processed new feature graph is used as the input of the CNN model to judge the category of abnormal sounds by training the CNN model. The results of measured data show that the combination of CNN and mixed features can accurately judge and classify abnormal sounds. And compared with existing CNN models trained directly using short-time Fourier time-frequency map or single MFCC feature image, the method proposed in this paper has higher recognition accuracy and faster training speed, and has good practical use value. Section 2 describes the realization process and principle of the feature graph, section 3 explains the structure of the CNN model, section 4 gives the experimental results and analysis, and section 5 summarizes.

2. Feature Extraction

In abnormal sound recognition, MFCC feature can solve the problem of insufficient ability of multiplicative combined sound signal to show difference, so it has research advantages over other audio features. This paper also tries to combine MFCC features with other audio features to improve the recognition performance. In this paper, short-term energy features and MFCC features are selected for fusion.

2.1. Mel Frequency Cepstral Coefficient

The extraction of MEL Frequency Cepstral coefficient (MFCC) is shown in figure 1, mainly including the following parts. Audio input and preprocessing. The power spectrum is obtained by computing the discrete cosine transform (DCT) or the fast Fourier transform (FFT) of each frame. The energy spectrum is obtained by modulating the power spectrum. The Mel spectrum is obtained by energy spectrum passing through the Mel filter bank. After taking the logarithm of Mel spectrum and performing DCT, Mel spectrum was obtained. Finally, the appropriate MFCC was selected.

This paper uses the method of windowing and framing to extract MFCC features, where each 1024 points is a frame, and the sliding distance between frames is set to 512 points, that is, the two frames before and after are overlapped. Hamming window is selected for adding Windows, and FFT transformation is selected for each frame. The filter bank consists of 60 Mel filters, and finally extracts 40-dimensional MFCC features.

2.2. Short-Term Energy

The short-term energy reflects the strength of the signal at different moments, and it also has a good effect on distinguishing the voiced sounds. In view of the presence of human voice and animal voice in the abnormal sounds in this paper, so it is taken as a part of the mixed feature. Let the audio time-
The domain signal be $x(n)$, the windowing function is $w(n)$, and the i-th frame audio signal obtained after frame processing is $y_i(n)$, then $y_i(n)$ satisfies equation (1).

$$\begin{align*}
\left\{ \begin{array}{l}
y_i(n) = w(n) \ast x((i-1) \ast inc + n) \\
1 \leq n \leq L, 1 \leq i \leq f_n
\end{array} \right.
\end{align*}
$$

In the formula, $w(n)$ is a window function, generally a rectangular window or a Hamming window, $y_i(n)$ is a frame value, $n = 1, 2, \ldots L$, $i = 1, 2, \ldots f_n$, $L$ is Frame length, inc is the frame shift length, and $f_n$ is the total number of frames after framing. Then, the short-term energy formula of the i-th frame speech signal $y_i(n)$ is as shown in equation (2).

$$E(i) = \sum_{n=0}^{L-1} y_i^2(n), 1 \leq i \leq f_n$$

For extracting short-term energy features, keep the frame setting when extracting MFCC features, calculate the energy of each frame according to equations (1) and (2), and finally obtain the one-dimensional feature.

Matrix stitching of 40-dimensional MFCC and 1-dimensional energy features yields 41-dimensional audio feature image. The feature image is normalized to obtain a new audio feature image.

3. Convolutional Neural Network

CNN is a type of deep learning network model designed to handle two-dimensional feature image recognition. It can directly take two-dimensional image as input and can realize feature learning and expression through layer-by-layer feature transfer. It requires less preprocessing and strong learning ability, and has been widely used in handwritten character recognition [8], face recognition [9] and other fields and has achieved good results. It differs from the standard neural network in that it replaces the fully connected hidden layer of the standard neural network with the combination of multiple convolution layer and sampling layer.

The convolution layer convolves the output of the previous layer with the convolution kernel, then adds the bias vector of the layer, and obtains the output of the layer through the nonlinear excitation function. In this process, each convolution kernel shares the same parameters, which essentially reduces the number of parameters of CNN. Meanwhile, the selection of convolution kernel also needs to balance the local correlation of features and the detailed information contained, which directly affects the quality of feature extraction [10]. The sampling layer is usually behind the convolution layer. According to some sampling rules, the feature image of the output of the convolution layer is dimensioned. The sampling methods generally include maximum sampling and mean sampling. In the
The sampling layer guarantees the scale invariance of the feature to some extent. Finally, a full connection layer is generally added after the last sampling layer, mainly for the purpose of integrating all features into one-dimensional information and facilitating classification.

The structure of the CNN model used in this paper is shown in figure 2. The model consists of an input layer, hidden layers, fully connected layers, and an output layer. Where, the hidden layer consists of two convolution layers C1, C2 and two sampling layers S1 and S2 alternately. The number of convolution kernels of the convolution layer C1 and C2 is 32 and 64, and the size is 3*3 and 2*2, respectively. The activation function is “Relu”, and the boundary can be filled with 0, so the size of the image remains unchanged after convolution operation. “MaxPooling” method is adopted in the sampling layer, the area size is 2*2, and the area does not overlap. The full connection layer and the output layer constitute the classifier, and the “Softmax” classifier is used to output the judgment results of 6 kinds of abnormal sounds.

![Model structure diagram of CNN.](image)

**Figure 2.** Model structure diagram of CNN.

4. Experimental Process and Results

4.1. Data Collection

A total of 9,000 pieces of abnormal sound data, for example gunshots and explosions, were collected in this article, which are mainly from various authorized audio databases and sound effects websites on the Internet, as well as manual collection by members of this experiment. The audio database contains UrbanSound8K [11], DCASE2016 [12], NIGENS, AudioSet, and the sound effects website contains BBC Sound FX [13], FREESOUND, ADOBE. In addition, this experiment collected part of dog bark and human scream by manual recording, and mixed them into the existing data set after noise elimination. In this experiment, all audio data in the data set is resampled, the sampling frequency is 44.1kHz, the quantization accuracy is 16bit, and the encoding is PCM. Thus, the sample information of the audio data set of this experiment is shown in table 1.

| Class   | Quantity | Training | Verification | Test  | Label                |
|---------|----------|----------|--------------|-------|----------------------|
| Explosion     | 900      | 300      | 300          | [1,0,0,0,0]T |
| Alarm          | 900      | 300      | 300          | [0,1,0,0,0]T |
| Glass Crash    | 900      | 300      | 300          | [0,0,1,0,0]T |
| Scream         | 900      | 300      | 300          | [0,0,0,1,0]T |
| Gun Shot       | 900      | 300      | 300          | [0,0,0,0,1]T |
| Dog Bark       | 900      | 300      | 300          | [0,0,0,0,1]T |
| Total          | 5400     | 1800     | 1800         |       |                      |
The characteristic image of glass broken sound after feature extraction is shown in figure 3. The left side is the simple MFCC characteristic image, and the right side is the mixed characteristic image after adding short-term energy. It can be seen that the newly added one-dimensional energy feature can well represent the energy intensity in the signal without any adverse effect on the MFCC feature.

Figure 3. Feature image of glass breaking.

4.2. Model Training Results

This experiment uses Tensorflow, which is a very versatile deep learning framework that can relatively easily create very complex neural network architectures [14]. In addition, this experiment uses a dynamic time warping method to normalize all the abnormal sound samples to the same number of frames to ensure that the characteristic parameters of each abnormal sound sample have the same size, to ensure the convenience of inputting the CNN model. For training, the model uses a micro-batch stochastic gradient descent method to optimize the cross-entropy loss function. Each batch contains 32 data randomly selected from the training set, and the set learning rate is 0.001. To cope with the overfitting phenomenon, the model also uses Dropout () to randomly disconnect 50% of the neurons after S1 and S2.

The experimental results are shown in figures 4-6. The training results obtained by simply using MFCC feature as the input of CNN network are shown in figure 4, and there is no overfitting in the whole process. After 128 steps of training, the accuracy was 0.938. The CNN training results using the mixed feature graph are shown in figure 5. After the same 128 steps of training, the accuracy is 0.962 and the loss function is stable below 0.2. In comparison, the results of CNN training with mixed feature graph are better than those with MFCC feature alone. The CNN training results directly using time-frequency graph as input are shown in figure 6. After 128 steps of training, the accuracy reaches 0.968, and the loss function is also controlled below 0.2. However, because the time-frequency map contains too much information, the training results are easy to repeat, and the response in the figure is that the accuracy and loss function curve jitter are relatively large. By comparing figures 5 and 6, it can be seen that the training results with mixed features as input are basically the same as the training results directly using time-frequency graph as feature input. In addition, the training speed of network is much faster than that of time-frequency graph.
Figure 4. Training results of MFCC.

Figure 5. Training results of mixed features.

Figure 6. Training results of time-spectrum graph.

5. Conclusion
In this paper, a noiseless abnormal sound data set is created, and CNN is applied to abnormal sound detection. This paper uses the MFCC and short-term energy features of abnormal sounds to propose a new mixed feature image, which is used as the input of CNN to obtain more accurate detection results. The experimental data show that in the field of audio classification and recognition, the combination of CNN and mixed features can give accurate judgment and classification for abnormal sound recognition. And compared with the existing direct short-time Fourier time-frequency map, the proposed method has faster training speed and excellent recognition accuracy. Compared with a CNN model trained with a single MFCC feature image, the proposed method has higher recognition
accuracy. So, this method has good practical application value. However, the effect of this method on the classification and recognition of noisy abnormal sounds is unknown, so how to use CNN to classify and recognize noisy abnormal sounds will be the future research direction of this paper.

References
[1] Schruder J, Anemuller J and Goetze S 2016 Performance comparison of GMM, HMM and DNN based approaches for acoustic event detection within Task 3 of the DCASE2016 challenge Proc. Workshop Detect. Classification Acoustic (Scenes Events) pp 80-84.
[2] Zhou X, Zhuang X, Liu M, Tang H, Hasegawa-Johnson M and Huang T 2017 HMM-based acoustic event detection with ada boost feature selection in classification of events Activeties and Relationships Evaluation and Workshop pp 345-353.
[3] Mesaros A, Heittola T, Dikmen O and Virtanen T 2015 Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity an-notations International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp 151-155.
[4] Valenzise G, Gerosa L, Tagliasacchi M, et al. 2007 Scream And Gunshot detection and localization for audio surveillance systems IEEE Conference on Advanced Video and Signal Based Surveillance pp 21-26.
[5] Cakir E, Parascandolo G, Heittola T, et al. 2017 Convolutional recurrent neural networks for polyphonic sound event detection IEEE/ACM Transactions on Audio, Speech, and Language Processing 25 (6) 1291-1303.
[6] Piczak K J 2015 Environmental sound classification with convolutional neural networks IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP) pp 1-6.
[7] Abdel-Hamid O, Mohamed A, Jiang H, et al. 2012 Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp 4277-4280.
[8] Youwang W 2014 Deep Learning and Its Application in Handwritten Chinese Character Recognition (South China University of Technology).
[9] Meng Y L, Shan X L and Shi G S 2015 AU-inspired deep networks for facial expression feature learning Neurocomputing 159 126-136.
[10] Qingqing Z, Yong L, Jielin P, et al. 2015 Continuous speech recognition based on convolutional neural network Journal of Engineering Science 37 (9) 1212-1217.
[11] Salamon J, Jacoby C and Bello J 2014 A dataset and taxonomy for urban sound research in Proc. ACM International Conference on Multimedia pp 1041-1044.
[12] Mesaros A, Heittola T, and Virtanen T 2016 TUT database for acoustic scene classification and sound event detection Proc. 24th European Signal Processing Conference pp 1128-1132.
[13] Slaney M 2002 Semantic-audio retrieval Proc. of International Conference on Acoustics, Speech and Signal Processing pp 1408-1411.
[14] Abadi M, et al. 2016 Tensorflow: Large-scale machine learning on heterogeneous distributed systems arXiv preprint arXiv:1603.04467.