ACOUSTIC SCENE CLASSIFICATION BASED ON 2-ORDER DENSE CONVOLUTIONAL NETWORK

Technical Report

Hongbo Fei[1,2], *Zilong Huang*[1,2], *Chen Liu*[1,2], *Yi Cao*[1,2,3*]

[1]School of Mechanical Engineering, Jiangnan University, Wuxi 214122, Jiangsu, China
[2]Jiangsu Key Laboratory of Advanced Food Manufacturing Equipment and Technology, Wuxi 214122, Jiangsu, China
[3]caoyi@jiangnan.edu.cn

ABSTRACT

In this technical report, we describe our acoustic scene classification algorithm submitted in DCASE 2020 Task 1a. We focus on network innovation, a novel acoustic scene classification model based on 2-order dense convolutional network is proposed, which aims at the problems of insufficient classification accuracy and adaptability of current models. Based on the dense convolutional neural network, combined with the N-order Markov model, the traditional dense connection is improved to the N-order correlation connection, and then the N-order dense convolutional network model is proposed. In terms of audio feature extraction, we use Log-Mel spectrograms and Gamma-Tone spectrograms to stitch together. In order to further improve system performance, virtual data generation technology is adopted. Finally, use the trained model for transfer learning. By using proposed systems, we achieved a classification accuracy of 69.16% on the officially provided evaluation dataset, which is 15.06% over than the baseline system.

Index Terms— Acoustic scene classification, 2-order dense convolutional network, N-order Markov model, Log-Mel spectrograms, Gamma-Tone spectrograms

1. INTRODUCTION

Audio carry a large amount of life scenes and physical events in the city [1], which plays an important role in our life. From the perspective of human cognition, auditory cognition is an important part of artificial intelligence. In the study of cognitive science, auditory cognition is often regarded as the second perception system second only to vision. Obviously, auditory cognition, as an important way to perceive the environment, its research value and development potential are self-evident. Acoustic scene classification(ASC) aims to classify sounds into one of predefined classes [2]. The audio scene classification competition and related conferences are also in full swing with the development of ASC. The DCASE Challenge was organized and launched by the University of London Queen Mary College Digital Music Center and Tampere University of Technology in 2013. It is currently the most authoritative competition in the field of acoustic events. Since 2016, the DCASE Challenge has been accompanied by a seminar which is held once a year, and many experts and scholars participate in it every year.

A high-quality dataset is an important prerequisite for testing whether the sound scene classification system is excellent. The DCASE challenge releases new dataset every year. From the DCASE 2016 challenge to the DCASE 2017 challenge, the length of each audio sample has been reduced from 30s to 10s [3]. The dataset released by the DCASE 2018 Challenge records high-quality binaural audio from 6 European cities as samples [4]. The dataset released by the DCASE 2019 Challenge ensures that each audio sample is recorded by the same device. With the release of the DCASE 2020 challenge, audio samples recorded by multiple devices and scenes are added to the data set for the first time, thereby further improving the quality of the data set.

In this report, we introduce an acoustic scene classification model based on N-order dense convolutional neural network. It is based on the dense convolutional neural network and combined with the improvement of the N-order Markov model, which is a more powerful dense convolutional neural network model. As for audio feature extraction, dual feature stitching is used, and after feature stitching, virtual data generation technology is used to achieve the purpose of data enhancement.

The remainder of this report is organized as follows. Section 2 describes the data preprocessing scheme. Section 3 details the structure and principles of N-order dense convolutional neural network in detail. Section 4 introduces the experiment and its results. Section 5 sets out the final conclusion.

2. DATA PREPROCESSION

This section describes our method of converting audio samples into acoustic features, and the method of data enhancement by generating virtual samples after feature stitching is completed.

2.1. Acoustic Feature

The sampling rate of the audio samples is the original 22050Hz, the length of the Fourier change window is set to 1024-samples (23ms) and the frame-shift is set to 1026- samples, and then each audio sample is divided into 215 frames.

2.1.1 Log-Mel Spectrogram

Using the short-time Fourier transform to obtain the spectrogram

after the audio samples that have been framed, and then pass through the 128-bit Mel filter bank, and finally take the logarithmic processing to obtain (N, 215, 128) shape Log-Mel spectrogram features [5].

$$mel(f) = \frac{1000}{log\,2} log\,(1 + \frac{f}{1000}) \quad (1)$$

2.1.2 Gamma-Tone Spectrogram

The method of extracting the Gamma-Tone spectrogram is similar to the above method, except that the Mel filter bank in the above method is replaced with the Gamma-Tone filter banks based on the ERB scale. The Gamma-Tone filter bank is a cochlear standard filter. It is a filter bank that simulates the human ear auditory system. The classic model of the filter bank impulse response is given as:

$$g_i(t) = At^{N-1}\, exp(-2\pi ERB(f_i)t) \times cos(2\pi f_i t + \varphi_i)\, U(t),$$
$$t \geq 0, 1 \leq i \leq N \quad (2)$$

where A is the filter gain, N is the filter order, f_i is the center frequency, φ_i is the phase, after simplifying the model, $\varphi_i = 0$. $ERB(f_i)$ is equivalent rectangular bandwidth. It determines the attenuation rate of the impulse response, which is related to the filter bandwidth, and each filter bandwidth is related to the critical frequency band of the human ear hearing [6]. In auditory psychology:

$$ERB(f) = 24.7 \times (4.37\frac{f}{1000} + 1) \quad (3)$$

to obtain the (N, 215,128) shape of the Gamma-Tone spectrogram features [7].

Finally, the two acoustic features are stitched together to obtain a synthetic feature vector of shape dimension (N, 215, 256).

2.2. Data Augmentation

In recent years, the number of samples in the dataset released by the DACSE challenge has increased year by year, but in order to improve the generalization ability of the model, it is not enough to just use the given samples. Therefore, the use of data augmentation to generate additional virtual data has gradually become the mainstream.

The mixup method is a form of neighborhood risk minimization [8]. This is an unconventional data enhancement method. Its principle is to extract additional virtual samples from the neighborhood distribution of training samples to expand the support of the sample distribution. The training distribution is expanded by fusing linear interpolation of feature vectors.

Mixup augmented data is obtained as follows:

$$\tilde{x} = \lambda x_i + (1 - \lambda)x_j \quad (4)$$

$$\tilde{y} = \lambda y_i + (1 - \lambda)y_j \quad (5)$$

where (x_i, y_i) and (x_j, y_j) are two acoustic scenes randomly chosen from the training data and $\lambda \in (0,1)$[9]. λ is acquired from the beta distribution and $\beta \in (0.1,0.9)$

3. NETWORK FRAMEWORK

In this part, firstly we introduce a new N-order dense convolutional network model (N-DenseNet) that we propose, and then explain the principles of forward propagation and back propagation of this novel network model through a 2-order dense convolutional network (2-DenseNet). The 2-order sub-model of N-DenseNet network will be used as the network model in the experiment in the next section.

3.1. N-Order dense convolutional network

It is obviously that dense convolutional network model (DenseNet) [10] provides the necessary theoretical basis for the proposal of N-DenseNet, therefore, the main principle of DenseNet would be concisely introduced here. The input of each layer in DenseNet comes from the output of all previous layers. Dense Block is a basic unit of DenseNet. In a l-layer Dense Block structure, the input of each layer is defined as $x_0, x_1, x_2, \ldots, x_l$, then:

$$x_l = H([x_0, x_1, \ldots, x_{l-1}]) \quad (6)$$

where $[x_0, x_1, \ldots, x_{l-1}]$ refers to concatenation of the feature-map produced in layers 0, …,$l-1$. $H(\cdot)$ represents the composite function of three consecutive operations: batch normalization (BN), followed by a rectified linear unit (RELU) and a 3×3 convolution (Conv).

Markov model is a statistical model widely used in speech recognition, natural language processing and other fields. The transition of its current state depends on the discrete time stochastic process of the first N states and has nothing to do with the previous historical state, which is called the N-order Markov model [11].

Combining characteristics of the DenseNet and N-order Markov model, N-DenseNet is designed as a novel network architecture to further improve the classification accuracy and generalization ability in ASC. N-Dense Block as a basic module of N-DenseNet, its input of l layer is just referred to the output of last N layers. In a l-layer N-Dense Block structure, the input of each layer is defined as $x_0, x_1, x_2, \ldots, x_{l-N}, \ldots, x_{l-1}, x_l$, then:

$$x_l = H([x_{l-N}, x_{l-N+1}, \ldots, x_{l-1}, x_l]) \quad (7)$$

In order to well complete the DCASE 2020 task 1a, 2-DenseNet as a sub-model of N-DenseNet is designed and shows better classification accuracy and generalization ability. Therefore, the 2-DenseNet and its working principle will be interpreted at length in the followings.

The state-dependent connection of 2-Dense block can be defined by the fact that the input of the l layer is just only related to the output of the previous 2 layers. In a l-layer 2-Dense Block structure, the input of each layer is defined as $x_0, x_1, x_2, \ldots, x_l$, the forward propagation of l-layer can be defined as:

$$\boldsymbol{X}^l(i,j) = [\boldsymbol{X}^{l-1} \otimes \boldsymbol{w}](i,j) + \boldsymbol{b}$$
$$= \sum_k \sum_m \sum_n \left[\boldsymbol{X}_k^{l-1}(i+m, j+n)\boldsymbol{w}_k(x,y)\right] + \boldsymbol{b} \quad (8)$$

where $\sum$ denotes the forward propagation of the convolutional layer, $\boldsymbol{X}^{l-1}$ and $\boldsymbol{X}^l$ represent the input and output of the feature-map, $\otimes$ denotes the convolution operation, $\boldsymbol{w}$ denotes the kernel function and $\boldsymbol{b}$ denotes the offset value, $X(i,j)$

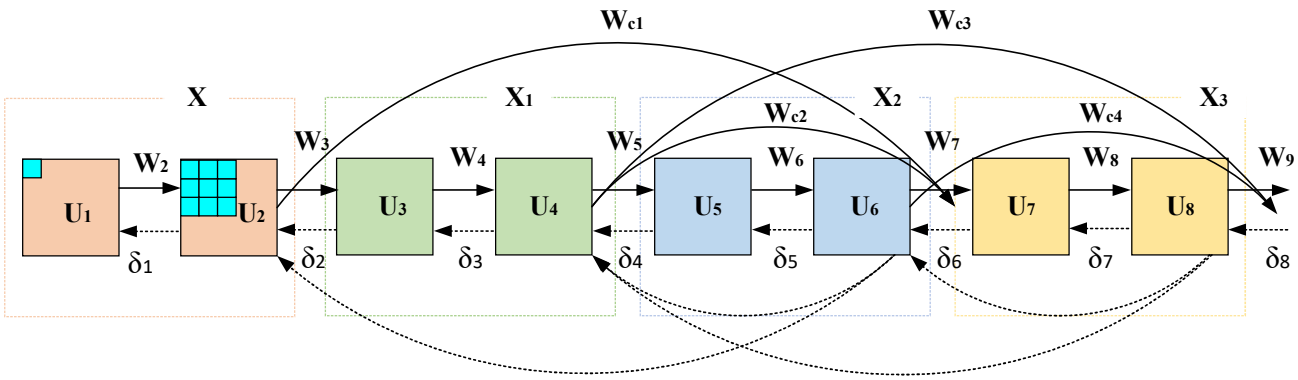

Figure 1: Schematic diagram of the forward propagation and back propagation of the 2-DenseNet

represents the pixel on the feature map, k is the number of feature map channels, m and n are the size of the convolution kernel.

For a 2-order state-dependent connection, the current layer in a 2-DenseBlock model is the concatenation layer by targeted and regulated tailoring of concatenation, that is, the input of the current layer just comes from the output of the previous 2-layers, which can be defined by:

$$x_l = H([x_{l-2}, x_{l-1}, x_l]) \tag{9}$$

3.2. Forward propagation and back propagation of 2-Densenet

In a l-layers Dense block structure, the number of state-dependent connection of DenseNet is $l(l-1)/2$, while in the 2-DenseNet, the number of state-dependent connection is $2(l-2)$. Actually, reducing the number of state-dependent connection leads to a faster convergence speed and a higher training efficiency of the network model. In this part, the principle of forward propagation and back propagation of 2-DenseNet will be introduced separately.

3.2.1 Forward propagation of 2-Densenet

The forward propagation schematic diagram of 2-DenseNet is shown in Figure 1.

As the basic convolution unit, a convolution combination structure such as 1×1 convolution and 3×3 convolution is shown in Figure 1 $\boldsymbol{X}$. The input of each layer is defined by: $\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_l}$, the feature-map output $\boldsymbol{U_i}$ of each layer starting from third layer is defined by:

$$U_i = f(BN(W_{3\times3} \otimes f(BN(W_{1\times1} \otimes [X_{l-2}, X_{l-1}, X_l] + B)))) \tag{10}$$

where $[X_{l-2}, X_{l-1}, X_l]$ represents the related connection mode of 2-Dense block, and uses the feature mapping of two previous layers as inputs, $W_{1\times1}$ and $W_{3\times3}$ indicates that the convolution kernel size is 1×1 and 3×3, $BN(\cdot)$ is the batch normalization, and $f(\cdot)$ is the activation function of RELU.

It is obviously that the related connection mode of 2-Dense block ensures the convergence speed of training process.

3.2.2 Back propagation of 2-Densenet

Back propagation is based on weight update strategy, which adjusts parameters on the negative gradient direction of target. It means the weight of each layer is continuously updated during the Back propagation. The method of weight update of 2-DenseNet is shown in Table 1.

In Table 1 and Figure 1, L represents the loss function of 2-DenseNet, δ_l represents the error of every layer, U_l indicates the output of each layer, while W_l is the matrix of the convolution layer and W_{cl} is the matrix of layer after $H(\cdot)$ operation, $*$ is the convolution operation to flip.

Table 1 shows the error term of 2-order concatenation layer is propagated back to the two previous layers, which can be expressed by:

$$\partial L/\partial W_{l-2} = \left(\delta_{l-1} * W_{l-1} + \delta_l * W_{cl} + \delta_{l+1} * W_{c(l+1)}\right) \otimes X_{l-2} \tag{11}$$

Table 1: Back Propagation parameters of 2-DenseNet

Layer	Back Propagation	Layer	Back Propagation
Input	$\partial L/\partial W_1 = \delta_1 * W_1 \otimes X$	L_5	$\delta_5 = \delta_6 * W_6 \otimes (\partial U_6/\partial U_5)$
L_1	$\delta_1 = \delta_2 * W_2 \otimes (\partial U_2/\partial U_1)$	L_6	$\delta_6 = \delta_7 * W_7 + \delta_8 * W_{c4}$
L_2	$\delta_2 = \delta_3 * W_3 + \delta_6 * W_{c1}$	L_7	$\delta_7 = \delta_8 * W_8 \otimes (\partial U_8/\partial U_7)$
L_3	$\delta_3 = \delta_4 * W_4 \otimes (\partial U_4/\partial U_3)$	L_8	$\delta_8 = \delta_9 * W_9 \otimes (\partial U_9/\partial U_8)$
L_4	$\delta_4 = \delta_5 * W_5 + \delta_6 * W_{c2} + \delta_8 * W_{c3}$	Output	$\partial L/\partial U_9$

4. EXPERIMENTS AND RESULTS

4.1. Datasets

The dataset for this task is TAU Urban Acoustic Scenes 2020 Mobile. The dataset contains recordings from 12 European cities in 10 different acoustic scenes using 4 different devices. Additionally, synthetic data for 11 mobile devices was created based on the original recordings. Of the 12 cities, two are present only in the evaluation set.

4.2. Training strategy

We use the officially provided fold 1 procedure to evaluate our systems' performance. Then the systems are retrained on the whole development data for submission. The train set is split into the train and evaluation set. The classifiers were trained on the trainset in maximum 400 epochs. Based on 2-DenseNet model, which described in Table 2, for a given number of convolution layers (24), convolution kernel size, channel number (k). The growth rate for all the network is k=40, optimizer Adam, Batch size = 32, etc.

Table 2: The architecture of 2-DenseNet

Layer	Description
Convolution(1)	5×5 conv-32-BN-RELU
Pooling	2×3 average pool
2-Dense Block (1)	$\begin{bmatrix}1\times1\text{ conv}\\3\times3\text{ conv}\end{bmatrix}\times3$
Transition Layer (1)	1×1 conv 2×2 max pool
2-Dense Block (2)	$\begin{bmatrix}1\times1\text{ conv}\\3\times3\text{ conv}\end{bmatrix}\times3$
Transition Layer (2)	1×1 conv 2×2 max pool
2-Dense Block (3)	$\begin{bmatrix}1\times1\text{ conv}\\3\times3\text{ conv}\end{bmatrix}\times3$
Transition Layer (3)	1×1 conv 2×2 max pool
Convolution(2)	1×1 conv-256
Pooling	Global Average pool
Dense	Dense (256, activation='relu')
Dense	Dense (10, activation='softmax')

4.3. Result

Results of experiments of various acoustics features on the fold 1 evaluation set is described in Table 3.

Analyzing the experimental results in Table 3, it can be found that the accuracy of the method of single acoustic feature extracting such as Gamma-Tone spectrograms feature extracting and Log-Mel spectrograms feature extracting is lower than the method of feature stitching process.

The accuracy of the method of Gamma-Tone spectrograms feature and the accuracy of the method of Log-Mel spectrograms feature extracting show that Single audio feature extraction method has certain limitations, which is not conducive to the improvement of audio classification tasks. While the dual feature stitching technology makes up for this shortcoming, which combines the advantages of two spectrograms and gets a better classification accuracy.

In ensemble system, Log-Mel spectrograms and Gamma-Tone spectrograms feature and stitched feature could be relatively complemented under a combination strategy. Table 3 shows the result of our algorithms.

In order to express the classification ability of our model, the confusion matrix of the 2-DenseNet is shown in Figure 2.

Table 3: The results of experiments

Method	Classification accuracy (%)
Baseline	54.10
Gamma-Tone spectrograms (GT-spec)	60.54
Log-Mel spectrograms (LM-spec)	61.32
GT-spec ⊕ LM-spec	64.44
Ensemble (proposed system)	69.16

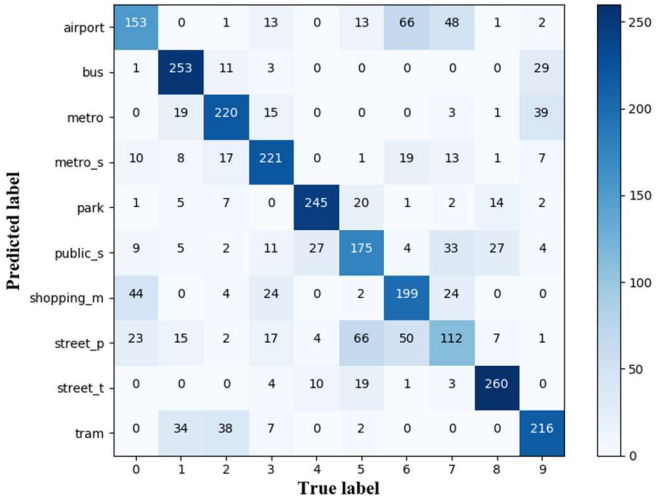

Figure 2: Confusion matrix of the ensemble system with 69.16% accuracy.

5. CONCLUSIONS

In this technical report, we proposed a novel network acoustic scene classification model based on 2-order dense convolutional network. Combined with audio features stitching technology, the best reliable result achieves 64.44% at the time when this technical report is submitted. After ensemble system, the final system accuracy rate can reach 69.16%, which is 15.06% over than the baseline system.

6. REFERENCES

[1] Z. Huang, C. Liu, and H. Fei, et al. Urban sound classification based on 2-order dense convolutional network using dual features [J]. Applied Acoustics 2020, 164.

[2] A. Mesaros, T. Heittola, and T. Virtanen, "Tut database for acoustic scene classification and sound event detection," pp.1128–1132, 2016.

[3] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," IEEE AASP Challenge on DCASE 2018 Technical Report, 2018.

[4] H. Yang, C. Shi, H. Li, "Acoustic scene classification using CNN ensembles and primary ambient extraction," IEEE AASP Challenge on DCASE 2019 Technical Report, 2019.

[5] T. Virtanen, M. D. Plumbley, D. Ellis, Computational Analysis of Sound Scenes and Events. [M] p. 76–78.

[6] Y. Wang, Z. Qian, X. Wang, et al. "An Auditory Feature Extraction Algorithm Based on γ-tone Filter-Banks" [J]. Acta Electronica Sinica 2010,03-0525-04

[7] B. R. Glasberg, B. C. Moore. Derivation of auditory filter shapes from notched noise data. Hear Res 1990; 47(1–2):103–38.

[8] H. Zhang, M. Cisse, Y. N. Dauphin,and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," in arXiv: 1710.09412, 2017.

[9] H Seo, J Park, Y Park, "Acoustic scene classification using various pre-processed features and convolution neural networks," IEEE AASP Challenge on DCASE 2019 Technical Report, 2019.

[10] H Gao, Z Liu, K. Q. Weinberger, "Densely Connected Convolutional Networks". In: 30th IEEE conference on computer vision and pattern recognition CVPR 2017, Honolulu, USA.p.2261-9

[11] J. Munkhammar, J. Widén. An N-state Markov-chain Mixture Distribution Model of The Clear-sky Index [J]. Solar Energy,2018,173(1):487-495.