# ACOUSTIC SCENE CLASSIFICATION USING MULTI-SCALE FEATURES

*Liping Yang , Xinxing Chen, Lianjie Tao*

Key Laboratory of Optoelectronic Technology and System(Chongqing University), Ministry of Education,
Chongqing University, China
{yanglp, gun_xing, taolianjie007} @cqu.edu.cn

## ABSTRACT

Convolutional neural networks(CNNs) has shown tremendous ability in many classification problems, because it could improve classification performance by extracting abstract features. In this paper, we use CNNs to calculate features layer by layer. With the layers deepen, the extracted features become more abstract, but the shallow features are also very useful for classification. So we propose a method that fuses features of different layers(it's called multi-scale features), which can improve performance of acoustic scene classification. In our method, the logMel features of audio signal are used as the input of CNNs. In order to reduce the parameters' number, we use Xception as the foundation network, which is a CNNs with depthwise separable convolution operation (a depthwise convolution followed by a pointwise convolution). And we modify Xception to fuse multi-scale features. We also introduce the focal loss, to further improve classification performance. This method can achieve commendable result, whether the audio recordings are collected by same device(subtask A) or by different devices (subtask B).

***Index Terms***— Multi-scale features, acoustic scene classification, convolutional neural network, Xception, logMel features

## 1. INTRODUCTION

Acoustic scene classification is a very complex problem which aim is to recognize the surrounding environment using acoustic signals. It has been used in many filed, such as context-aware services [1], surveillance [2] and robotic navigation [3]. Acoustic scene classification is so important that it has been attracting the attention of researchers in machine learning communities. The consecutive editions of the IEEE AASP Challenges Detection and Classification of Acoustic Scenes and Events(DCASE) [4] release the open and established datasets, and provide the scenario to evaluate and benchmark different approaches for acoustic scene classification and acoustic event detection, which makes the research of acoustic scene classification develop at full speed. Nowadays, many methods have been applied to acoustic scene classification, such as signal processing and machine learning, including dictionary learning [5], matrix factorization [6][7], wavelet filterbanks [8], and recently popular deep learning, such as CNN [9], Gated Recurrent Neural Networks(GRNN) [10].

The general framework for acoustic scene classification usually contains two steps. Fist obtain 2D time-frequency representation of data, and extracting relevant features. Second employ these features to achieve classification. And the most commonly used in acoustic scene classification is the Mel Frequency Cepstral Coefficients (MFCC) [3] and logMel features [18]. Different from a short-time Fourier transform(STFT), the constant Q transformation (CQT) provides a frequency analysis on a log-scale which makes it more adapted to sound and music representations, so the spectrum based on the CQT is also used in acoustic scene classification [19]. After computing the 2D time-frequency representation, some methods have investigated many features that are typically used in computer vision such as histogram of gradients (HOG) [19] and local binary pattern (LBP) [20]. Recently, some researchers have proposed feature learning, and learning features from spectrograms can provide representations that are adapted to the data while addressing the general lack of flexibility of hand-crafted features [6][7].

More recently, methods based on Deep Neural Networks (DNNs) have achieved good performance for acoustic scene classification. In [9], the authors presented a CNNs architecture with localized (small) kernels for environmental sound classification, and proposed data augmentation to overcome the problem of data scarcity. In [21], authors presented a distributed sensor server system for acoustic scene classification in urban environment based on CNNs. To exploit sequential correlation and local spectrum-temporal information, some researchers combined the long short term memory units (LSTM) and CNNs in parallel as lower networks [22].

In this paper, we present a new acoustic scene classification method. We fuse the multi-scale features to improve performance of acoustic scene classification. In order to reduce the number of parameters, we use Xception as the foundation network [11], which is convolutional neural network entirely based on depthwise separable convolution layers, and the Xception architecture is a linear stack of depthwise separable convolution layers with residual connections [11]. We modify the Xception architecture, via taking the output of last three blocks, and global pooling the output of each block. Then concatenate them together to achieve multi-scale feature fusion. The output of each block characterize different features, and the deeper blocks have more abstract features. Considering that features of each block have effect on the acoustic scene classification, we fuse the output of these block, and use multi-scale features to improve classification performance. We also introduce the focal loss [13] to further improve classification performance. Our method can achieve good results on subtask A and subtask B.

The rest of this paper is organized as follows. Section 2 presents modified Xception for acoustic scene classification, and describe how to perform multi-scale feature fusion. Section 3 discusses our experiments and results. Section 4 concludes our work.

## 2. PROPOSED METHOD

This section introduces the proposed multi-scale features fusion, modified Xception and focal loss of multi-class classification.

### 2.1. Multi-scale features

CNNs have powerful feature extraction capabilities, which realizes feature extraction and dimensionality reduction through operations such as convolution and pooling. The previous classification methods only used the last feature map, the feature map followed by some fully-connected layers(FC), the FC not only has a large amount of parameters, but also has a large amount of calculation. So at present, the most common methods usually perform global pooling on the last feature map, and then use the softmax layer to achieve classification [12][15].

In image classification, using the last feature map's information only, can achieve great performance. But in our case, its performance is not satisfactory. In the field of object detection, some researchers have used multi-scale feature maps to improve detection performance[16]. Inspired by this idea, we use multi-scale features to improve classification performance. The different feature maps contain different information, these information is helpful for improving the classification performance. In Fig. 1, we illustrated how to fuse the multi-scale features. We use the last three feature maps, and perform global pooling on the features, then concatenate them for fusion.

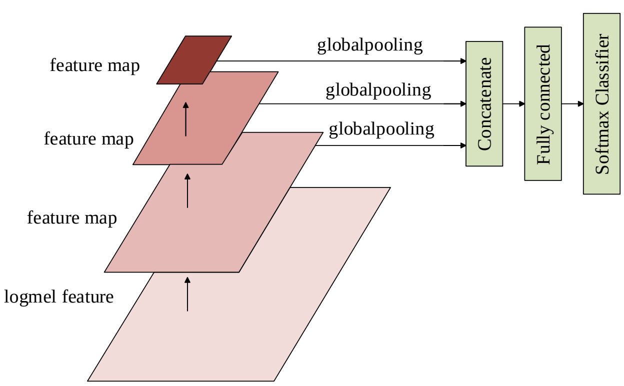


Figure 1: Illustrate how to fuse the multi-scale features

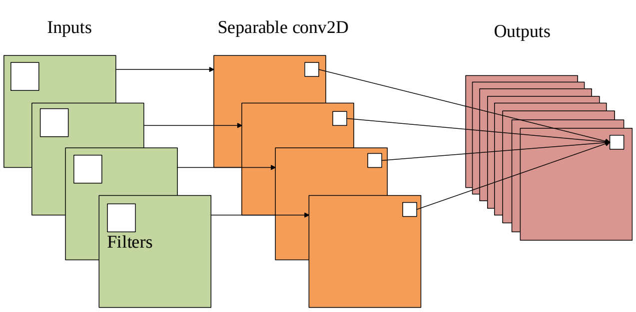


Figure 2: A block of depthwise separable convolution

### 2.2. Modified Xception

Xception is a convolutional neural network architecture entirely based on depthwise separable convolution layers[12]. In depthwise separable convolution, the convolution operation is split into multiple steps, as shown in Fig. 2. To better illustrate the depthwise separable convolution, we suppose that is a $3\times3$ size convolutional layer with a 16 channels input and a 32 channels output. The general convolution uses 32 convolution kernels

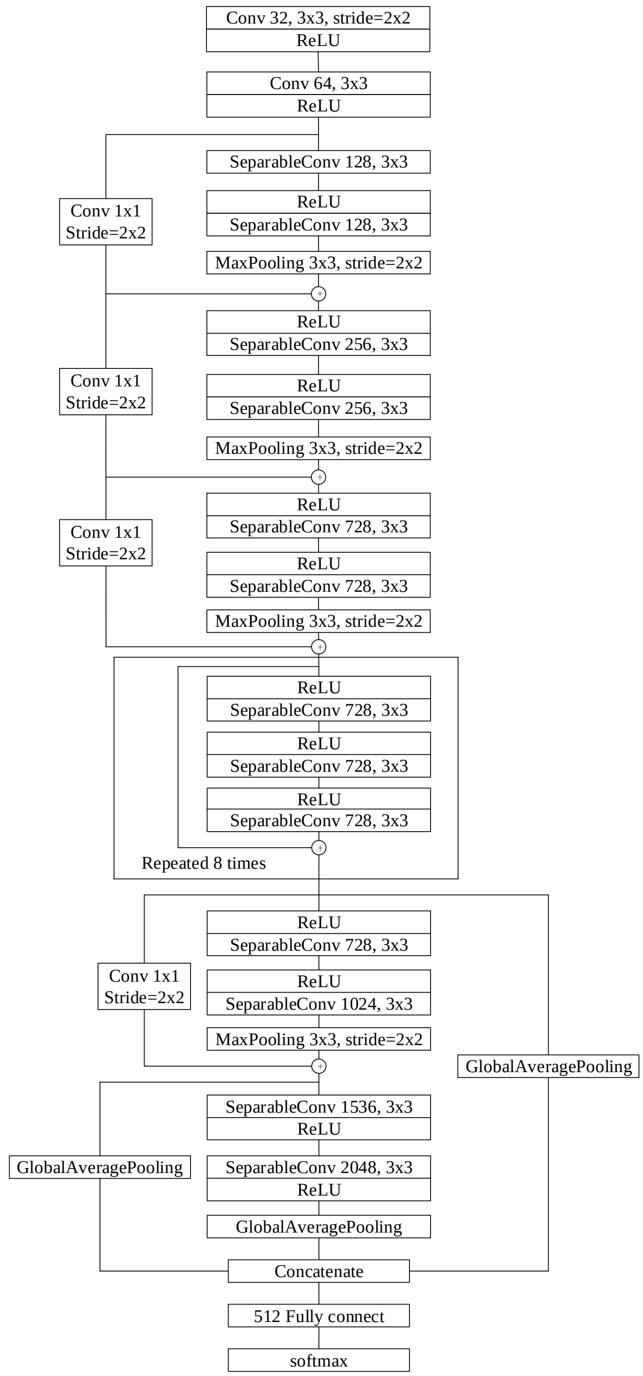


Figure 3: Overview of the modified Xception

convolving with input data, in which $3\times3\times16$ parameters are needed for each convolution kernel. And the output is only one channel. Then 32 convolution kernels need a total of $(3\times3\times16)\times32 = 4068$ parameters.

Depthwise separable convolution is split two steps. First, *depthwise convolution*, which is a spatial convolution performed independently over each channel of one input, 16 convolution kernels (1 channel) of $3\times3$ size are convoluted with 16 channels input data respectively. Second, *pointwise convolution*, which is a 1x1 (16 channels) convolution, projecting the 32 channels output by the depthwise convolution onto a new channel space. These two steps need $3\times3\times16+(1\times1\times16)\times32 = 656$ parameters, which has less amount of parameters than ordinary convolution. And depthwise separable convolutions are usually implemented without non-linearities activation function.

A complete description of the specifications of the network is given in Fig. 3. The Xception architecture has 36 convolutional layers forming the feature extraction base of the network. The 36 convolutional layers are structured into 14 modules, all of which have linear residual connections around them, except for the first and last modules. We extract the feature maps of the 32nd, 34th, and 36th layers, and perform global pooling on features maps respectively, then concatenate the outputs of global pooling. We fuse the features through FC layer and use softmax layer to perform classification.

### 2.3. Focal loss

For multi-class classification task, Cross-Entropy (CE) is generally used as the loss function:

$$CE(p,y) = -\sum_{j=0}^{c} y_j \log(p_j) \tag{1}$$

where, $p$ is the model's estimated probability, $y$ is ground-truth class label(one-hot vector), $j$ represents the $j$-th class. In this paper, we use loss function of acoustic scene classification that is based on CE:

$$l(p,y) = -\frac{1}{n}\sum_{i=0}^{n}\sum_{j=1}^{c} y_j^i \log(p_j^i + \varepsilon) \tag{2}$$

where, $y^i$ represents the label of $i$-th sample. $p^i$ represents the predicted label of $i$-th sample, $j$ represents the $j$-th class. $\varepsilon$ is a small positive number to prevent the occurrence of 0 in the logarithmic function.

During the training process, we found, some samples are hard to recognition. These samples would affect the prediction performance of our model. Therefore, we introduce the focal loss [13], the original focal loss start from the CE loss for binary classification. In this paper, we need a multi-class classification loss function, therefore we modify the focal loss. First, we define the probability $p_t^i$ that the $i$-th sample is predicted correctly:

$$p_t^i = (y^i)^T * p^i \tag{3}$$

where, $(y^i)^T$ represents the transpose of the $i$-th sample's label, $p^i$ represent the predicted label of $i$-th sample, $*$ is vector multiplication. The finally loss function are as follows:

$$L(p,y) = -\frac{1}{n}\sum_{i=0}^{n}(1-p_t^i+\varepsilon)^{\gamma}\log(p_t^i+\varepsilon) \tag{4}$$

The modified focal loss, can solve the problem of hard recognition samples, and we only need to select the appropriate hyper parameter $\gamma$. Known by definition of the focal loss, for hard to recognize sample, its probability $p_t^i$ is close to 0, and the $(1-p_t^i+\varepsilon)^{\gamma}$ is large. For easy to recognize sample, its probability $p_t^i$ is close to 1, $(1-p_t^i+\varepsilon)^{\gamma}$ is small, so it can down-weight loss of easy sample and up-weight loss of hard sample. It focus training on hard sample.

## 3. EXPERIMENTAL

### 3.1. Experimental Setting

We perform experiments on the dataset of DCASE2018 Task1 Subtask A and Subtask B, which consists of 10 scenes, *airport, shopping mall, metro station, street pedestrian, public square, street traffic, tram, bus, metro* and *park*. We use test set and train set divided by DCASE2018 committee.

Log-scaled mel-spectrogram are used as the input representation of the network. To compute it, the 2-channel wav of subtask A are down mixed to mono, and the wav of subtask B are mono. And STFT is applied using Hamming windows of 4096 samples with 75% overlap. After calculating its power, a mel filter bank is applied consisting of 128 bands. Then we use a filter bank with triangular filters in the frequency domain presenting a peak value of one. Finally, the resulting mel energy values are logarithmically scaled. Resulting log-scaled mel-spectrograms are normalized to zero mean and unit standard deviation for the training set.

The network training was performed by optimizing the focal loss and stochastic gradient descent (SGD) with Nesterov momentum. In the focal loss, $\gamma=3$ for subtask A, and $\gamma=1$ for subtask B, $\gamma$ is the optimal values selected by hyper-parameter search. The initial learning rate, and mini-batch size were set to 0.1, and 128, respectively, and use automatic attenuation of learning rate. We train network on dataset for 100 epochs, if the performance of the model is improved after training one epoch, the weight of the model is saved, if the performance of model is not improved after continuous 5 epochs, the learning rate is multiplied by 0.1, and if the performance of model is not improved after continuous 15 epochs, we stop training model.

### 3.2. Comparison with baselines

Our first experiment compares our method to baseline, the baseline system implements a CNNs based approach, where 40 log mel-band energies are first extracted for each 10-second signal, and a network consisting of two CNNs layers and one fully connected layer is trained to assign scene labels to the audio signals [17]. we perform experiments on development datasets of subtask A and subtask B.

Table 2 presents the results of our proposed method and baseline system. Compared with the baseline system, our proposed method achieves a relative improvement of more than 20%, on subtask A and subtask B.

Table 1: Comparing performances of baseline and our method on the subtask A and subtask B.

| Scene | Accuracy(%) Baseline | | Our method | |
|---|---|---|---|---|
| | Subtask A | Subtask B | Subtask A | Subtask B |
| Airport | 72.9 | 73.3 | **77.3** | **78.1** |
| Bus | 62.9 | 59.4 | **84.4** | **88.7** |
| Metro | 51.2 | 43.3 | **79.3** | **72.4** |
| Metro station | 55.4 | 50.4 | **86.8** | **87.8** |
| Park | 79.1 | 78.1 | **86.9** | **91.0** |
| Public square | 40.4 | 36.2 | **51.2** | **53.1** |
| Shopping mall | 49.6 | 48.2 | **88.7** | **79.7** |
| Street, pedestrian | 50.0 | 51.1 | **76.7** | **62.9** |
| Street, traffic | 80.5 | 80.5 | **91.2** | **87.5** |
| Tram | 55.1 | 51.9 | **75.0** | **74.6** |
| **Average** | 59.7 | 57.2 | **79.8** | **77.6** |

Table 2: Analyzing the effects of multi-scale features on the subtask A and subtask B, *w/o* means not using multi-scale features, and *with* means using multi-scale features. In this experiment we don't use the focal loss.

| Scene | Accuracy(%) w/o multi-scale features | | with multi-scale features | |
|---|---|---|---|---|
| | Subtask A | Subtask B | Subtask A | Subtask B |
| Airport | **78.5** | 76.4 | 77.1 | **77.8** |
| Bus | **88.4** | 81.7 | 84.9 | **89.0** |
| Metro | 74.3 | **73.7** | **78.9** | 71.2 |
| Metro station | 85.7 | 85.1 | **87.1** | **87.8** |
| Park | **88.8** | 90.7 | 86.7 | **91.2** |
| Public square | **53.7** | 48.0 | 47.3 | **49.8** |
| Shopping mall | 72.7 | 75.6 | **88.6** | **79.2** |
| Street, pedestrian | 65.2 | **63.3** | **75.4** | 61.4 |
| Street, traffic | 86.6 | 86.2 | **92.3** | **87.6** |
| Tram | 74.1 | 73.1 | **74.2** | **74.3** |
| **Average** | 76.8 | 75.3 | **79.3** | **76.9** |

### 3.3. On the effect of multi-scale features

Our second experiment analyze the effect of multi-scale features on performance. In this experiment, we don't use focal loss, and perform it on development datasets of subtask A and subtask B.

Table 2 presents the results of our proposed method with multi-scale features and without multi-scale features. On subtask A, the method with multi-scale features achieves 2.5% relative improvement compared with the method without multi-scale features, and on subtask B, the improvement is 1.6%. It can been seen that fusion of multi-scale features can improve performance.

### 3.4. On the effect of focal loss

Our third experiment analyze the effect of focal loss, In this experiment, we use multi-scale features. And perform this experiment on development datasets of subtask A and subtask B.

Table 2 presents the results of our proposed method with focal loss and without focal loss. The focal loss could solve the problem that some samples are difficult to recognize, the method with focal loss achieves 0.6% improvement on subtask A, and 0.7% improvement on subtask B.

Through these experiments, we can draw conclusions, our method can achieve great classification performance on subtask A and subtask B.

Table 3: Analyzing the effects of the focal loss on the subtask A and subtask B, *w/o* means not using focal loss, and *with* means using focal loss. In this experiment, we use multi-scale features.

| Scene | Accuracy(%) w/o focal loss | | with focal loss | |
|---|---|---|---|---|
| | Subtask A | Subtask B | Subtask A | Subtask B |
| Airport | 77.1 | 77.8 | **77.3** | **78.1** |
| Bus | **84.9** | **89.0** | 84.4 | 88.7 |
| Metro | 78.9 | 71.2 | **79.3** | **72.4** |
| Metro station | **87.1** | **87.8** | 86.8 | **87.8** |
| Park | 86.7 | **91.2** | **86.9** | 91.0 |
| Public square | 47.3 | 49.8 | **51.2** | **53.1** |
| Shopping mall | 88.6 | 79.2 | **88.7** | **79.7** |
| Street, pedestrian | 75.4 | 61.4 | **76.7** | **62.9** |
| Street, traffic | **92.3** | **87.6** | 91.2 | 87.5 |
| Tram | 74.2 | 74.3 | **75.0** | **74.6** |
| **Average** | 79.3 | 76.9 | **79.8** | **77.6** |

## 4. CONCLUSION

In this paper, we propose an acoustic scene classification method which uses multi-scale features fusion. We use Xception as the foundation network, in order to fuse features, we modify the Xception. This method can achieve great classification performance on subtask A and subtask B. In order to further improve performance, we introduce focal loss of multi-class classification. Although our method is still satisfactory, its biggest problem is the existence of overfitting, and if there are more data to train our model, we would get better performance.

## 5. REFERENCES


[1] A. J. Eronen et al., “Audio-based context recognition,” *IEEE Trans. Audio,Speech, Lang. Process.*, vol. 14, no. 1, pp. 321–329, Jan. 2006.

[2] R. Radhakrishnan, A. Divakaran, and P. Smaragdis, “Audio analysis for surveillance applications,” in *Proc. IEEE Workshop Appl. Signal Process.Audio Acoust.*, 2005, pp. 158-161.

[3] S. Chu, S. Narayanan, C.-C. J. Kuo, and M. J. Mataric, “Where am I? Scene recognition for mobile robots using audio features,” in *Proc. IEEE Int. Conf. Multimedia Expo*, 2006, pp. 885–888.

[4] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah,E. Vincent, B. Raj, and T. Virtanen, “DCASE 2017 challenge setup: Tasks, datasets and baseline system,” in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, submitted.

[5] J. Salamon and J. P. Bello, “Unsupervised feature learning for urban sound classification,” in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 2015, pp. 171–175.

[6] V. Bisot, R. Serizel, S. Essid, and G. Richard, “Acoustic scene classification with matrix factorization for unsupervised feature learning,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICA SSP)*, Shanghai, China, Mar. 2016, pp. 6445–6449.

[7] V. Bisot, R. Serizel, S. Essid, and G. Richard, “Feature learning with matrix factorization applied to acoustic scene classification,” in *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 6, pp. 1216-1228. June 2017

[8] J. T. Geiger and K. Helwani, “Improving event detection for audio surveillance using gabor filterbank features,” in *23rd European Signal Processing Conference (EUSIPCO)*, Nice, France, Aug. 2015, pp. 714–718.

[9] J. Salamon, J. P. Bello, “Deep convolutional neural networks and data augmentation for environmental sound classification,” *IEEE Signal Processing Letters*, 2017, pp (99):1-1.

[10] Z. Ren, V. Pandit, K. Qian, et al, “Deep sequential image features for acoustic scene classification,” in *Proc. of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, pp. 113-117

[11] F. Chollet, "Xception: deep learning with depthwise separable convolutions." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1800-1807.

[12] K. He, X. Zhang, S. Ren, and J. Sun. “Deep residual learning for image recognition,” *arXiv preprint arXiv:1512. 03385*, 2015.

[13] T. Y. Lin, P. Goyal, R. Girshick, et al, “Focal loss for dense object detection,” In *Proceedings of the IEEE Conf-erence on Computer Vision and Pattern Recognition*, 2017, pp. 2999-3007.

[14] G. Huang, Z. Liu, et al, “Densely connected convolutional networks,” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2261-2269.

[15] K. Simonyan, A. Zisserman. “Very deep convolutional networks for large-scale image recognition,” In *ICLR*, 2015.

[16] W. Liu, D. Anguelov, et al, “SSD: single shot multibox detector,” in *European Conference on Computer Vision*. Springer, Cham, 2016, pp. 21-37.

[17] http://dcase.community/challenge2018/task-acoustic-scene-classification .

[18] S. Park, S. Mun, Y. Lee, et al, “Acoustic scene classification based on convolutional neural network using double image features,” in *Proc. of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, pp. 98-102.

[19] A. Rakotomamonjy, G. Gasso, “Histogram of gradients of time-frequency representations for audio scene classification,” *IEEE Trans. Audio,Speech, Lang. Process.*, vol. 23, no.1, pp. 142-153, Jan. 2015.

[20] W. Yang, S. Krishnan, “Combining temporal features by local binary pattern for acoustic scene classification,” *IEEE Trans. Audio,Speech, Lang. Process.*, vol. 25, no. 6, pp.131 5-1324, June 2017.

[21] J. Abeßer, S. I. Mimilakis, et al. “Acoustic scene classification by combining autoencoder-based dimensionality reduction and convolutional neural networks,” in *Proc. of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, pp.7-11.

[22] S. H. Bae, I. Choi, and N. S. Kim, “Acoustic scene classification using parallel combination of LSTM and CNN,” *Proc. of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, Sept-ember 2016, pp. 11-15