Acoustic Scene Classification Based on Dense Convolutional Networks Incorporating Multi-channel Features

To cite this article: Dezhi Wang et al 2019 J. Phys.: Conf. Ser. 1169 012037

View the article online for updates and enhancements.
Acoustic Scene Classification Based on Dense Convolutional Networks Incorporating Multi-channel Features

Dezhi Wang¹, Lilun Zhang, Kele Xu² and Yongxian Wang¹
¹College of Meteorology and Oceanography, National University of Defense Technology, Sanyi Street, Changsha, 410073, China
²College of Information and Communication, National University of Defense Technology, Wuhan, 410073, China
*E-mail: wang_dezhi@hotmail.com

Abstract. Motivated by the state-of-the-art performance of Dense Convolutional Networks (DenseNet) on highly competitive object recognition benchmark tasks (CIFAR-10, CIFAR-100, SVHN and ImageNet), this work presents improvements to the Acoustic Scene Classification (ASC) task of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE 2017) based on an optimized DenseNet model. Multi-channel Convolutional Neural Network (CNN) is also explored to extract features from different audio channels in an end-to-end manner. In the experiments, the proposed model is compared with the challenge baseline model of DCASE 2017 and several other state-of-the-art CNN architectures subject to the classification accuracy on the same open-source DCASE datasets. The results show that the proposed DenseNet-based architecture can achieve a superior performance in classification accuracy but with a lower model complexity in contrast with other models.

1. Introduction
Implementing an automated system that understands acoustic scenes is promising, since a variety of information can be acquired from an audio segment. Acoustic Scene Classification (ASC) in the DCASE challenges [1] is a task of characterizing the surrounding environment under a predefined set of scene classes based on given audio segments. Various methods have been explored in ASC research. For instance, using log mel-band energies as features, Gaussian Mixture Model (GMM) [2] and Multilayer Perceptron architecture (MLP) are included in the baseline model in DCASE 2017 competition [1].

Deep learning based approaches are concentrated in this study since deep neural networks have become a state-of-the-art system for ASC, thanks to the recent advances in deep learning research [3] especially in the fields of image recognition [4] and speech recognition [5]. In particular, the CNN networks can combine hierarchical feature extraction and classification together. In the field of acoustic signal processing, CNN-based models show a great improvement in performance as opposed to most traditional classifiers. Several popular CNN structures have been proposed sequentially, such as AlexNet [4], VGG [6], GoogLeNet (Inception) [7], and ResNet [8]. Moreover, CNNs are also designed for audio recognition tasks such as speech recognition [9], music transcription [10], and environmental sound classification [11]. CNNs are normally used to extract ‘deep features’ from the spectrograms of segmented audio waves for ASC task. Since CNNs are more robust and stable if
being trained on a large scale data set of (audio) samples, such nets can be reused in a pre-trained manner from other large image datasets for the targeting ASC problems through transfer learning [12]. However, transfer learning is not allowed in the DCASE 2017 challege since it takes advantage of extra dataset.

Recently, a Dense Convolutional Network (DenseNet) is developed and has yielded significant improvements in accuracy for object recognition tasks [13]. DenseNet employs direct connections between each convolutional layer to every other layer in a feed-forward fashion, which allows convolutional networks to be substantially deeper (an alternative way is with the help of skip connections [8]), more accurate, and more efficient to train. Due to direct connections, DenseNet requires fewer parameters than traditional convolutional networks (like recent variations of ResNets [14]) due to the fact that redundant feature-maps do not need to be re-learned. Also the flow of information and gradients throughout the network has been improved for a fast training convergence [15]. DenseNet has achieved the state-of-the-art results across several highly competitive computer vision datasets under multiple settings, especially with fewer parameters and less computation. Moreover, DenseNet can be applied as good feature extractors because of its compact internal representations and reduced feature redundancy [16]. As far as we know, there is still no application of DenseNet in the ASC task, which motivates us to carry out this work.

Most of ASC approaches have worked on identifying acoustic scenes based on monaural audio signal. Different sound sources from different directions will have different intensities on channels. Multi-channel signals contain spatial information which could be employed to help to distinguish different sound sources. However, there is still little work using multi-channel information for the audio scene classification. The potential of combining multi-channel signals based on multi-channel CNN architectures for more detailed information of acoustic scenes is believed to lead to advanced feature representations for the classification and to improve the performance of classifiers. In view of this point, Multi-channel CNN architecture is also explored integrated with DenseNet in this study. A Multi-channel framework considers the interaural magnitude differences (IMD), which actually contains the spatial information of audios to improve the classification [17]. Audio time series normally contain the temporal and spectral structure features at different time and frequency scales [18], which requires a superior feature representations. Various time-frequency representations such as spectrograms are widely used in ASC tasks, which normally can offer a rich feature representation for classification.

The remaining part of this paper is organized as follows. Section 2 describes the architecture of the proposed model. Section 3 describes the experimental settings and configurations, along with the results and analysis. Section 4 give the conclusion of this paper.

2. Proposed method
The input for CNN architectures can be raw audio signals or feature sets extracted from the raw signals. The most widely used features are Mel-band energies [12], MFCCs, spectrograms and so on. In our study, mel-filterbank energies of the audio signal segments are employed as the feature representations to be fed into the proposed DenseNet-based system. It is also worth noting that it is easy to extend our framework for other types of input features.

Unlike the attempts which aim to maintain the one-channel CNN architecture [14], the combination of information in multi-channels may lead to advanced feature representations for the classification. In our experiments, features are extracted from three different parts including left channel, right channel, the difference between the left and right channels. In the experiments, the mel-filterbank energies are calculated for each channel with the same configuration where a Hann window of 1024 points is used with a hop size of 512 points. On the other hand, 224 mel-bands are used to generate the mel-filterbank energies while 224 frame-features are concatenated to form the input features used in model training. Thus the network input becomes images with size of 224×224×3, where 224×224 denotes the mel-filterbank feature size and 3 represents the number of channels.
As shown in Figure 1, the proposed system architecture comprises of dense convolutional blocks with direct connections from any layer to all subsequent layers to improve the information flow on 224×224 input images. Four dense blocks with unequal numbers of layers make up the DenseNet used in our experiments.

A convolution with 64 output channels is performed on the input images in front of the first dense block. For convolutional layers with kernel size 3×3, one-pixel padding is applied at each side of the inputs to keep the feature-map size fixed. The layers between two contiguous dense blocks are referred as transition layers for convolution and pooling, which contain 1×1 convolution and 2×2 average pooling. As shown in Table 1, a bottleneck layer is used by using a 1×1 convolution before each 3×3 convolution in order to reduce the number of input feature-maps and improve computational efficiency.

At the end of the last dense block, a global average pooling and a softmax classifier are employed. Table 1 illustrates the change of the feature-map sizes in each layer of the applied DenseNet architecture. In this study, we just select 3 commonly used DenseNet architectures i.e. DenseNet 121 with \([n_1,n_2,n_3,n_4] = [6,12,24,16]\), DenseNet 169 with \([6,12,32,32]\) and DenseNet 201 with \([6,12,48,32]\).

3. Experiments
In this section, details of the DCASE 2017 ASC dataset used in the study are first introduced. Then the model parameters and experimental settings are presented for the comparison of performance between the proposed DenseNet-based model and other the state-of-the-art CNN models. Finally, the experimental results are discussed.

3.1. Dataset
The 2017 DCASE challenge dataset for acoustic scene classification is used to validate the performance of proposed method. The dataset is established to determine the context of a given recording through selecting one appropriate label from a pre-determined set of 15 acoustic scenes such as cafe/restaurant, car, city center and so on. Each scene contains 312 recordings with a length of 10 seconds, a sampling rate of 44.1 kHz and 24-bit resolution in stereo in the development dataset. Totally there are 4680 audio recordings in the development dataset which is provided at the beginning of the challenge, together with ground truth. Besides an evaluation dataset is also released with 1620 audio recordings in total after the challenge submission is closed. A four-fold cross-validation setup is provided so as to make results reported strictly comparable. The evaluation dataset is used to finally evaluate the performance of classification models.

3.2. Model Comparison
In this study, we compare the performance of the proposed DenseNet-based system with four other popular CNN architectures, they are, VGG 19, ResNet 50, Inception V3 and Xception, which have achieved success in a lot of fields like computer vision. VGGNet is proposed in order to increase the depth of CNN network by reducing the size of receptive field. Multiple convolutional layers with small size 3×3 kernels are stacked on top of each other in the increasing depth. Unlike VGGNet, ResNet is instead a form of different architecture that relies on micro-architecture modules (also called network-in-network architectures). Moreover, Inception model is developed to overcome the inefficiency and high computational cost of VGG model by designing the multi-level feature extraction components which compute 1×1, 3×3, and 5×5 convolutions within the same module of the network. It should be noted that, as the newest model in the Inception group, Xception architecture employs the depth-wise separable convolution operation to replace the regular Inception modules, which has become a cornerstone of convolutional neural network architecture design. Thus, the Xception model is also employed in this study for a comparison.

In addition, the baseline system provided in the DCASE 2017 ASC task is also considered as a reference for comparison. The baseline system consists of 60 MFCC features and a Gaussian mixture
model based classifier. MFCC features are calculated using 40-ms frames with Hamming window and 50% overlap and 40 mel-bands. A GMM model with 32 components is trained for each scene class.

The DenseNet and other models are trained on the exactly same development dataset by using sub-segments of the audio recordings in terms of a four-fold cross-validation. The changes we make to the standard VGG 19, ResNet 50, Inception V3 and Xception are using the global average pooling layers to replace the original final layers and applying batch normalization instead of Local Response Normalization (LRN). Moreover, the last layer of all the models is the softmax output layer which outputs each probability corresponding to one class. The probability can be used to predict the scene label of the audio segment.

![DenseNet structure](image)

**Figure 1.** (Color online) A DenseNet-based network with four dense blocks. Inside a dense block each layer takes all preceding feature maps as input.

**Table 1.** DenseNet structure for ASC

| Layers               | Output Size | Settings                                      |
|----------------------|-------------|-----------------------------------------------|
| Convolution          | 112×112     | 3x3 conv, stride 2                            |
| Dense Block (1)      | 56×56       | \[
\begin{bmatrix}
1\times1 \text{conv} \\
3\times3 \text{conv}
\end{bmatrix} \times n_1, \text{growth rate 32}
\] |
| Transition Layer (1) | 56×56       | 1x1 conv, 2 x 2 average pool, stride 2        |
| Dense Block (2)      | 28×28       | \[
\begin{bmatrix}
1\times1 \text{conv} \\
3\times3 \text{conv}
\end{bmatrix} \times n_2, \text{growth rate 32}
\] |
| Transition Layer (2) | 28×28       | 1x1 conv, 2 x 2 average pool, stride 2        |
| Dense Block (3)      | 14×14       | \[
\begin{bmatrix}
1\times1 \text{conv} \\
3\times3 \text{conv}
\end{bmatrix} \times n_3, \text{growth rate 32}
\] |
| Transition Layer (3) | 14×14       | 1x1 conv, 2 x 2 average pool, stride 2        |
| Dense Block (4)      | 7×7         | \[
\begin{bmatrix}
1\times1 \text{conv} \\
3\times3 \text{conv}
\end{bmatrix} \times n_4, \text{growth rate 32}
\] |
| Prediction Layer     | 15×1        | ReLU activation, global average pooling 2D, softmax layer |
Each audio chunk is pre-processed by segmenting it using a 1024-point sliding window with a 512-point hop size. Each segment is converted into 224×224 dimensional mel-filterbank energies. The parameters of all the networks are generally tuned depending on the heuristic experience. All the deep-learning models are optimized with Adaptive Moment Estimation (ADAM) technique using a learning rate of 0.00003 with the aid of an early-stopping technique using a patience of 7. A batch size of 64 is applied in training all the models. The objective function is selected as the categorical-crossentropy. Moreover, an L2 weight decay penalty is selected in the training process. For performance evaluation, the averaged accuracy calculated based on the four-fold cross-validation dataset is used as the main metric.

3.3. Results
In order to fully utilize GPU resource, the DenseNet-based model and other models are implemented and trained on the basis of Keras with a GPU-version tensorflow as backend based on a hardware set of Xeon E5 2683V3 CPU and 2 GTX 1080Ti GPU cards, where the computation is also significantly accelerated by CUDA with cuDNN [19]. Table 2 illustrates the ASC accuracies obtained by different systems as well as the baseline system. All CNN-based models clearly outperform the baseline system on the given dataset in terms of ASC accuracy. The results also show that using DenseNet architecture generally performs better than using the VGG, ResNet, Inception and Xception. In addition, DenseNet models also have the smallest computational complexity according to the related running times (based on the training of 30 epoches) shown in Table 2.

Based on the results and our best knowledge, the proposed DenseNet-based architecture achieves the state-of-the-art performance on DCASE 2017 ASC dataset by using single end-to-end CNN model (rather than ensemble models). It may imply that the design of direct connections between every two convolutional layers can improve the performance of CNN model for ASC task.

| Method        | Depth | Parameters | Running Time in Hours (30 Epoches) | Average Accuracy (%) (Development Dataset) | Average Accuracy (%) (Evaluation Dataset) |
|---------------|-------|------------|-----------------------------------|---------------------------------------------|-------------------------------------------|
| DenseNet 121  | 40    | 7.05M      | 2.93                              | 77.1±2.19                                   | 65.2±2.56                                 |
| DenseNet 169  | 88    | 12.67M     | 4.64                              | 80.5±2.10                                   | 70.1±2.41                                 |
| **DenseNet 201** | **98** | **18.35M** | **5.13**                          | **81.3±1.91**                               | **71.5±2.01**                             |
| VGG 19        | 19    | 6.09M      | 5.77                              | 74.9±2.63                                   | 62.7±3.15                                 |
| Inception V3  | 177   | 21.833M    | 6.69                              | 76.4±2.23                                   | 66.3±2.66                                 |
| ResNet 50     | 100   | 13.618M    | 5.45                              | 80.1±2.79                                   | 69.5±2.88                                 |
| Xception      | 165   | 17.21M     | 4.93                              | 79.8±2.77                                   | 69.9±2.89                                 |
| DCASE 2017 ASC Baseline | -     | -          | -                                 | 74.8                                        | 61.0                                      |
4. Conclusions

Despite CNNs have shown superior performance than other methods in a variety of machine learning recognition and classification tasks, such approaches have not been widely explored in the domain of acoustic scene classification. In this paper, we propose a DenseNet-based system for ASC task in the DCASE 2017 challenge contest. Several state-of-the-art CNN network architectures are also considered on the same dataset for comparison. Thanks to the direct connections between convolutional layers, the optimized DenseNet architecture can achieve a superior performance in ASC accuracy but with a lower model complexity. Especially, the results show that our system is able to achieve an accuracy of 71.5% on the challenge evaluation dataset, i.e. an improvement of 10 percentage points over the baseline system. It should be noted that this accuracy is achieved based on a single deep-learning model rather than ensemble of a series of models. In future work, our system will be tested over a wide range of different acoustic classification tasks. In addition, the potential of DenseNet-based system will be investigated for acoustic deep representation learning.

Acknowledgements

This study was funded by the National Natural Science. Foundation of China (No. 61806214), the Scientific Research Project of NUDT (No.ZK17-03-31).

References

[1] Mesaros A, Heittola T, Diment A, Elizalde B, Shah A, Vincent E, Raj B, and Virtanen T. DCASE 2017 challenge setup: Tasks, datasets, and baseline system. in Detection and Classification of Acoustic Scenes and Events. 2017.
[2] Reynolds D A, Quatieri T F and Dunn R B. Speaker verification using adapted gaussian mixture models. in Digital Signal Processing. 2000.
[3] Yang T C I and Hsieh H. Classification of acoustic physiological signals based on deep learning neural networks with augmented features. in Computing in Cardiology Conference. 2017.
[4] Krizhevsky A, Sutskever I and Hinton G E, Imagenet classification with deep convolutional neural networks. Communications of the Acm, 2012. 60(2): p. 2012.
[5] Gong Y, Huang Y, Kumar K, Li J, Liu C, Ye G, Zhang S, Zhao Y, and Zhao R, Challenges in and solutions to deep learning network acoustic modeling in speech recognition products at microsoft. 2017.
[6] Ren Z, Pandit V, Qian K, Yang Z, Zhang Z, and Schuller B. Deep sequential image features for acoustic scene classification. in The Workshop on Detection & Classification of Acoustic Scenes & Events. 2017.
[7] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, and Rabinovich A. Going deeper with convolutions. in 2015 IEEE Computer Vision and Pattern Recognition (CVPR). 2015. Boston, United States.
[8] He K, Zhang X, Ren S and Sun J, Deep residual learning for image recognition, in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: Las Vegas, NV, United States. p. 770-778.
[9] Zhang Y, Pezeshki M, Brakel P, Zhang S, Laurent C, Bengio Y, and Courville A. Towards end-to-end speech recognition with deep convolutional neural networks. in the Annual Conference of the International Speech Communication Association (INTERSPEECH). 2016.
[10] Sigta S, Benetos E and Dixon S, An end-to-end neural network for polyphonic piano music transcription. 2016: IEEE Press. 927-939.
[11] Piczak K J. Environmental sound classification with convolutional neural networks. in IEEE International Workshop on Machine Learning for Signal Processing. 2015.
[12] Deng J, Cummins N, Han J, Xu X, Ren Z, Pandit V, Zhang Z, and Schuller B, The university of passau open emotion recognition system for the multimodal emotion challenge. CCPR 2016, Part II, CCIS 663, 2016: p. 652-666.
[13] Huang G, Liu Z, Weinberger K Q and Laurens V D M, Densely connected convolutional networks, in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: Honolulu, Hawaii, United States.

[14] Huang G, Sun Y, Liu Z, Sedra D, and Weinberger K Q. Deep networks with stochastic depth. in European Conference on Computer Vision. 2016.

[15] Lee C Y, Xie S, Gallagher P, Zhang Z, and Tu Z, Deeply-supervised nets. Eprint Arxiv, 2014: p. 562-570.

[16] Gatys L A, Ecker A S and Bethge M, A neural algorithm of artistic style. Computer Science, 2015.

[17] Xu Y, Kong Q, Huang Q, Wang W, and Plumbley M D, Convolutional gated recurrent neural network incorporating spatial features for audio tagging. arXiv:1702.07787, 2017.

[18] Cui Z, Chen W and Chen Y, Multi-scale convolutional neural networks for time series classification. arXiv:1603.06995, 2016.

[19] Serrano J, Nvidia introduces cudnn, a cuda-based library for deep neural networks. http://infoq.com, 2015.