Acoustic Scene Classification Using Deep Convolutional Neural Network via Transfer Learning

Min Ye[1,2], Hong Zhong[1], Xiao Song[2], Shilei Huang[3], Gang Cheng[2,*]
[1]School of Computer Science and Technology, Anhui University, Hefei, China
[2]PKU Shenzhen Institute, Shenzhen, China
[3]Shenzhen Raisound Technologies, Co., Ltd, Shenzhen, China
$\{min.ye, xiao.song\}$ @$imsl.org.cn$; $zhongh$@$ahu.edu.cn$; $\{shilei.huang, gang.cheng\}$ @$raisound.com$

We use deep convolutional neural network via transfer learning for Acoustic Scene Classification (ASC). For this purpose, a powerful and popular deep learning architecture — Residual Neural Network (Resnet) is adopted. Transfer learning is used to fine-tune the pre-trained Resnet model on the TUT Urban Acoustic Scenes 2018 dataset. Furthermore, the focal loss is used to improve overall performance. In order to reduce the chance of overfitting, data augmentation technique is applied based on mixup. Our best system has achieved an improvement of more than 10% in terms of class-wise accuracy with respect to the Detection and classification of acoustic scenes and events (DCASE) 2018 baseline system on the TUT Urban Acoustic Scenes 2018 dataset.

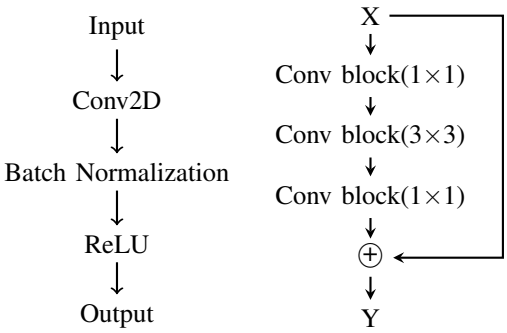
***Keywords*—transfer learning; Acoustic Scene Classification; focal loss; mixup**

I. Introduction

With the increase of the amount of data, the neural network with many layers shows superior performance. But if the network is deeper, the gradient exploding of the back propagation becomes a problem. ResNet [1] solved this problem and achieved state-of-the-art performance in the fields of image classification, object detection, instance segmentation, semantic segmentation. However, training Resnet from scratch requires large amounts of data and high computational resources. Transfer learning [2] save computational resources by applying knowledge learned from a problem to another different but related problem.

Acoustic Scene Classification (ASC) is the task to identify audio recordings that recorded in a public area into one of several predefined acoustic scene classes, such as "park", "pedestrian street" and "metro station". There have been several attempts to use transfer learning for ASC. For example, deep neural network based learning and transferring mid-level audio features for ASC [3], a study on transfer learning for Acoustic Event Detection (AED) in a real life scenario [4] and an investigation of transfer learning mechanism for ASC [5].

In this paper, we describe ASC system based on transfer learning for identifying audio that recorded in a public area into one of several predefined acoustic scene classes. For this purpose, we fine-tune pre-trained deep convolutional neural network of Resnet on the TUT Urban Acoustic Scenes 2018 dataset. We also adopt the focal loss [6] to further improve classification performance, data augmentation using the mixup [7] is introduced to diversify the training data.



(a) Bottleneck block (b) Residual block

Fig. 1. Detail of blocks.

The remaining content of this paper is organized as follows: Section II describes the transfer learning ASC system. Our experiments are described in Section III. Section IV is the conclusion.

II. Transfer learning ASC system

This section describes the transfer learning ASC system. The following subsections give more description on transfer learning and model fine-tuning, Resnet, focal loss and data augmentation using mixup.

A. Transfer learning and model fine-tuning

Transfer learning is a machine learning technique in which a model trained on one task is reused for another related task. If there is not enough data to train large networks from scratch, transfer learning can be used to avoid overfitting [8]. Transfer learning also saves computational resources.

In this paper, we train the pre-trained Resnet model and fine-tune the pre-training Resnet model weight [9] on the TUT Urban Scenes 2018 dataset. In this case, the number of nodes in the output layer must be modified to match the number of classes for ASC. In addition, the data must match the input size of the pre-training Resnet model.

B. Resnet

As we all know, network depth is the factor that determines network performance. Deeper and deeper networks are used in

978-1-7281-5014-7/19/$31.00

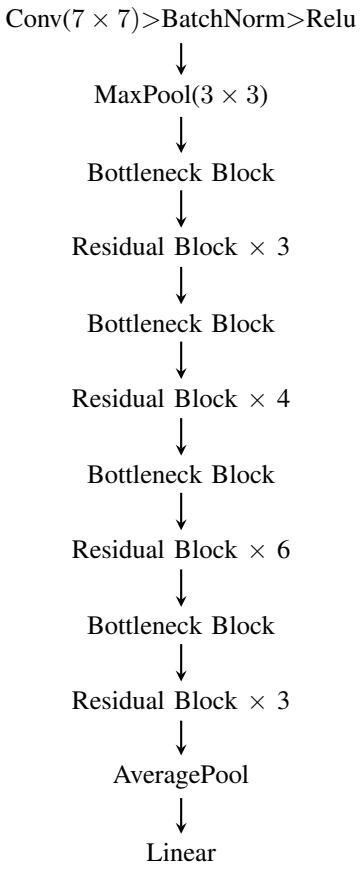


Fig. 2. Resnet architecture.

computer vision area. However, due to the gradient vanishing problem, deep network training is not easy. ResNet solved this problem and provide a training framework to ease the training of networks that are substantially deeper than those used previously.

The detail of blocks in this paper are shown in Fig. 1. Fig. 1 (a) shows the bottleneck block and Fig. 1 (b) shows the residual block. For bottleneck block, Batch Normalization (BN) [10] is added after convolution. The ReLU function is used as a non-linearity of Resnet after BN. Residual block have three bottleneck blocks. The convolution kernels of the three bottleneck blocks are 1×1, 3×3, 1×1, respectively. Residual block first reduces the input channel count with 1×1 kernels, applies 3×3 convolutions in that reduced channel size and then restores to the input channel size back with 1×1 convolutional layer.

The Resnet architecture in this paper are shown in Fig. 2. As show in Fig. 2, the layers of Residual block is 3, 4, 6, 3, respectively. Resnet is based on 50 convolutional layers in this paper.

C. Focal loss

We train the network with the extension of the focal loss [6], which was previously used for object detection in images of natural scenes using the stochastic gradient descent optimizer. Focal loss is an extension of Cross Entropy (CE) loss, which solves very large class imbalance problems and performs implicit negative mining by imposing higher losses on uncertain prediction. There may have the sample imbalance problem in ASC.

The normal CE loss for object detection is showed below:

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

where y specifies the ground-truth class and $p \in \{0, 1\}$ is the models estimated probability, j represents the j-th class. The CE can solve the problem that the weight update too slow. When the error is large, the weight update fast, so CE is widely used. The focal loss for object detection is showed below:

$$FL(p, y) = -\alpha_t (1 - p_t)^{\gamma} log(p_t) \tag{2}$$

In this paper, ASC is multiple classification but object detection is binary classification, therefore we modify the focal loss as follows:

$$Label(p, y) = -\frac{1}{n} \sum_{i=0}^{n} \alpha_t (1 - p_t^i)^{\gamma} log(p_t^i) \tag{3}$$

For focal loss, we only need to select the appropriate hyper-parameters α, γ.

D. Data Augmentation using Mixup

We uses mixup as a method of data augmentation. This method can improve the generalization ability of the model and construct a virtual training sample. The mathematical expression for the mixup is as follows:

$$\widehat{x} = \lambda x_i + (1 - \lambda) x_j \tag{4}$$

$$\widehat{y} = \lambda y_i + (1 - \lambda) y_j \tag{5}$$

Where (x_i,y_i) and (x_j,y_j) are two samples randomly extracted from the training data, and $\lambda \in [0, 1]$, $\lambda \sim Beta(\alpha, \alpha)$, $\alpha \in (0, \infty)$. This mixture extends the features of training set and label distribution through linear interpolation of eigenvectors and linear interpolation of corresponding labels. The super-parameter α of the mixup controls the interpolation strength of the features and labels.

III. Experimental

A. Datasets

TUT Urban Acoustic Scenes 2018 dataset [11] includes 10 categories of acoustic scenes such as “bus”, “airport”, “park” and “metro”. The dataset was recorded in six large European cities, and in different locations for each scene class. For each location there is one 5-6 minute audio recordings. The original recordings were split into segments with a length of 10 seconds that are provided in individual files. The dataset consists of 10-seconds audio segments from 10 acoustic scenes. TUT Urban Acoustic Scenes 2018 development dataset contains in total 24 hours of audio and each acoustic scene has 864 segments (144 minutes of audio). TUT Urban Acoustic Scenes 2018

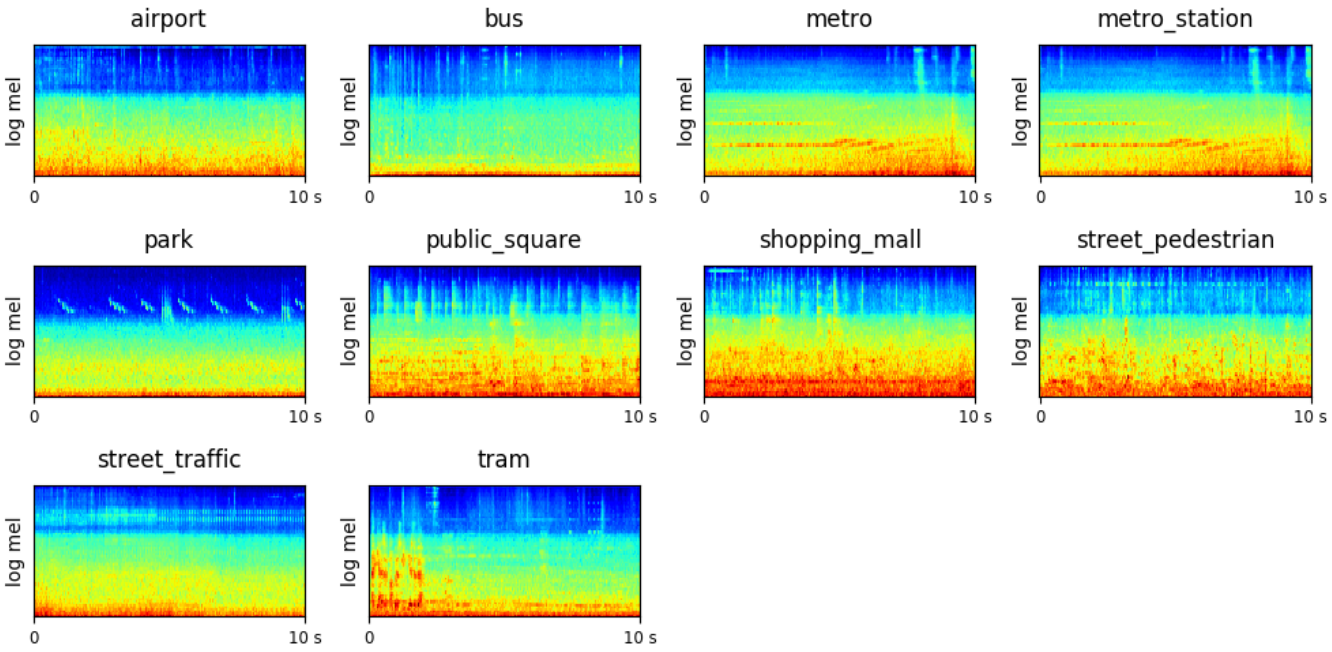


Fig. 3. The log-mel spectrograms of the scenes.

leaderboard dataset contains in total about 3.3 hours of audio and each acoustic scene has 120 segments (20 minutes of audio).

In this paper, we choose TUT Urban Acoustic Scenes 2018 development dataset as the training subset, which all contains 8,640 segments. Also, we choose TUT Urban Acoustic Scenes 2018 leaderboard dataset as the test subset, which all contains 1,200 segments.

B. Evaluation

The scoring of ASC will be based on classification accuracy: the number of correctly classified segments among the total number of segments. Accuracy will be calculated as average of the class-wise accuracy.

C. Feature extraction

The log-mel spectrograms of the scenes are shown in Fig. 3. We extract the spectrograms and apply log-mel filter banks on the spectrograms followed by logarithm operation where the sampling rate is set to be 48 kHz. A short-time Fourier transform (STFT) [12] with a Hanning window size of 2048 samples and a hop size of 1500 samples is used to extract the spectrogram, so that there are 320 frames in an audio clip. The mel filter bank has a cut-off frequency of 50 Hz. Mel filter banks with 64 bins and cut-off frequencies of 50 Hz to 24 kHz are applied on the spectrogram.

D. Setup

We use AdaBound [13] optimizer for gradient-based optimization [14] with learning rate of 0.001. The learning rate is reduced by multiplying 0.9 after every 200 iterations of training. As described in reference [13], AdaBound is an optimizer that behaves like Adam at the beginning of training, and gradually transforms to Stochastic Gradient Descent (SGD) at the end. The final learning rate parameter indicates AdaBound would transform to SGD with this learning rate. In common cases, a default final learning rate of 0.1 can achieve relatively good and stable results on unseen data. It is not very sensitive to its hyper-parameters. The transfer learning ASC system are optimized during 5000 maximum iteration steps, which are empirically set. The transfer learning ASC system are implemented with Python and Pytorch. The hyper-parameters α of mixup is 0.3.

TABLE I
CLASS-WISE ACCURACY OF THE ASC SYSTEM.

Model	Accuracy
Baseline [11]	0.625
SE-Resnet [15]	0.725
NLL loss	0.725
focal loss	0.733
focal loss + mixup	**0.747**

E. Results and discussion

TABLE I shows the class-wise accuracy of our systems on the TUT Urban Acoustic Scenes 2018 dataset. To confirm the performance of our ASC systems, we compared it with two conventional methods, The baseline [11] of the DCASE 2018 Challenge and SE-Resnet [15].

a) Compare our systems and baseline: The baseline of the DCASE 2018 Challenge is based on two convolutional layers. Compared with the 62.5% class-wise accuracy of the baseline system, the best class-wise accuracy achieved by our approach was 74.7% and improves class-wise accuracy by almost 10%.

b) Compare our systems and SE-Resnet: SE-Resnet is based on 152 convolutional layers of squeeze-and-excitation [16] Resnet. Our system is based on 50 convolutional layers of pre-trained Resnet. Compared with the 72.5% class-wise accuracy of the SE-Resnet, our best method achieves a relative improvement of 2.2% on the TUT Urban Acoustic Scenes 2018 dataset.

c) On the effect of focal loss: Our experiment analyze the effect of focal loss. NLL loss is CE loss. The class-wise accuracy of our system with NLL loss was 72.5% and the class-wise accuracy of our system with focal loss was 73.3%, respectively. The focal loss could solve the problem that some samples are difficult to recognize, the system with focal loss achieves 0.8% relative improvement for ASC.

d) On the effect of mixup: Our experiment analyze the effect of mixup. We use focal loss. The class-wise accuracy of our system without mixup was 73.3% and the class-wise accuracy of our system with mixup was 74.7%, respectively. The system with mixup achieves 1.4% relative improvement compared with the method without mixup.

IV. Conclusion

We applied transfer learning to identify acoustic scenes using deep convolutional neural network. The utilized network is based on pre-trained model of Resnet. We evaluated the network on the TUT Urban Acoustic Scenes 2018 dataset by fine-tuning the pre-trained model of Resnet. We use focal loss solve very large class imbalance problems and mixup as data augmentation. Our best model obtained more than 10% improvement over the baseline system of DCASE 2018 challenge on the TUT Urban Acoustic Scenes 2018 dataset.

Acknowledgment

This work was supported by Shenzhen Basic Research Program (JCYJ20170817155939233) and Shenzhen Technology Project (JSGG20170822105644555).

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[2] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, p. 9, 2016.

[3] S. Mun, S. Shon, W. Kim, D. K. Han, and H. Ko, "Deep neural network based learning and transferring mid-level audio features for acoustic scene classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 796–800.

[4] P. Arora and R. Haeb-Umbach, "A study on transfer learning for acoustic event detection in a real life scenario," in *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2017, pp. 1–6.

[5] H. Zhou, X. Bai, and J. Du, "An investigation of transfer learning mechanism for acoustic scene classification," in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2018, pp. 404–408.

[6] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[7] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[8] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in neural information processing systems*, 2014, pp. 3320–3328.

[9] O. Lucena, A. Junior, V. Moia, R. Souza, E. Valle, and R. Lotufo, "Transfer learning using convolutional neural networks for face anti-spoofing," in *International Conference Image Analysis and Recognition*. Springer, 2017, pp. 27–34.

[10] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[11] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," *arXiv preprint arXiv:1807.09840*, 2018.

[12] A. Nuruzzaman, O. Boyraz, and B. Jalali, "Time-stretched short-time fourier transform," *IEEE Transactions on Instrumentation and Measurement*, vol. 55, no. 2, pp. 598–602, 2006.

[13] L. Luo, Y. Xiong, Y. Liu, and X. Sun, "Adaptive gradient methods with dynamic bound of learning rate," in *Proceedings of the 7th International Conference on Learning Representations*, May 2019.

[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[15] J. H. Yang, N. K. Kim, and H. K. Kim, "Se-resnet with gan-based data augmentation applied to acoustic scene classification," DCASE2018 Challenge, Tech. Rep., September 2018.

[16] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.