An Investigation of Transfer Learning Mechanism for Acoustic Scene Classification

Hengshun Zhou[1]*, Xue Bai*[1]*, Jun Du*[1]

[1]National Engineering Laboratory of Speech and Language Information Processing, University of Science and Technology of China

zhhs@mail.ustc.edu.cn, byxue.mail.ustc.edu.cn,jundu@ustc.edu.cn

Abstract

One main challenge for acoustic scene classification (ASC) is there are remarkable overlaps and similarities between different acoustic scenes. However, the most existing ASC tasks are always lack of adequate training data to well distinguish different classes, especially in the deep learning approaches, such as using convolutional neural network (CNN). Motivated by the success of the transfer learning mechanism from the image classification task (e.g., ImageNet) with a large amount of training data to other computer vision tasks with less training data [1], in this study we investigate the possibility of transfer learning between two quite different classification tasks with the inputs of 2D image signals and 1D audio signals. One strong motivation behind this is the spectrograms of the audio signal can be also considered as the 2D images which are potentially have the similar structures to those samples in the image classification task. Specifically, we conduct the transfer learning mechanism by adopting the pre-trained CNNs with different architectures from the ImageNet task to the DCASE2018 ASC subtask A. Furthermore, by leveraging more input channels and training data fragments, the classification accuracy of our proposed system is increased from 59.7% to 77.8% on the evaluation set, in comparison to the officially provided CNN system trained using only audio data.

Index Terms: transfer learning, image classification, acoustic scene classification, convolutional neural network, pre-training

1. Introduction

Sounds carry a large amount of information regarding to the environment and physical events. Humans can perceive the sound scene (e.g., busy square, park), and recognize individual sound sources (e.g., bus passing by, footsteps). This process is called auditory scene analysis [2]. The research field studying this process is called computational auditory scene analysis (CASA) [3]. The computational algorithms attempt to automatically make sense of the environment through the analysis of sounds using signal processing and machine-learning methods. The corresponding task is called acoustic scene classification (ASC) [4], and the goal is to classify a test audio into one of predefined classes that characterizes the environment in which it was recorded, e.g., airport, park, metro station.

In the past few years, acoustic scene classification has been gradually receiving attention in the field of audio signal processing and machine learning. Substantial progress has been made by several important challenges, such as Detection and Classification of Acoustic Scenes and Events (DCASE2016) [5] and DCASE2017 [6]. Many new techniques have emerged and been widely investigated, including the aspects of feature deisigns, statistical models, decision criteria, and meta-algorithms. Several categories of audio features have been employed in acoustic scene classification systems, such as low-level time-based and frequency-based audio descriptors [7, 8], frequency-band energy features (energy/frequency) [7], auditory filter banks (Gammatone, Mel filters), Mel-Frequency Cepstral Coefficients (MFCCs), spatial features,like interaural time difference (ITD), interaural level difference (ILD) [9], voicing features (fundamental frequency f0) [10] and i-vector [11]. In DCASE2017, harmonic, percussive [12] and constant-Q transform (CQT) [13] have also been investigated, and achieved better results. The features described here can be further processed to derive new quantities that are used either in place or as an addition to the original features, like principal components analysis and time derivatives [14].

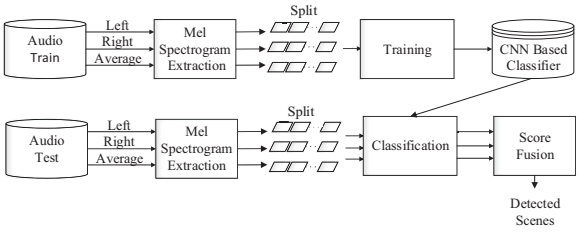


Figure 1: *System overview.*

Once the features are extracted from the audio samples, the next stage is learning statistical distribution models of the features. Statistical models can be divided into generative and discriminative methods. One classical generative model for acoustic scene classification is the Gaussian mixture models (GMM) [15] where features are interpreted as being generated by a sum of Gaussian distributions. MFCC features and maximum likelihood criterion are used for GMM training and testing. As for discriminative models, support vector machine (SVM) is a popular discriminative classifier for acoustic scene classification [9,]. In terms of training data augmentation, generative adversarial nets (GAN) scored first place in DCASE2017 [16]. Recently, convolutional neural networks (CNNs) have been investigated and applied much for music tagging [17], acoustic scene classification [18, 19]. CNN provides an effective way to capture spatial information of multidimensional data. And each feature map captures information at different locations in the picture.

One main problem for most acoustic scene classification tasks is the lack of training data to well distinguish the confusing and overlapping acoustic scenes. Motivated by the success of the transfer learning mechanism from the image classification task (e.g., ImageNet) with a large amount of training data to other computer vision tasks with less training data [1], in this study we investigate the possibility of transfer learning be-

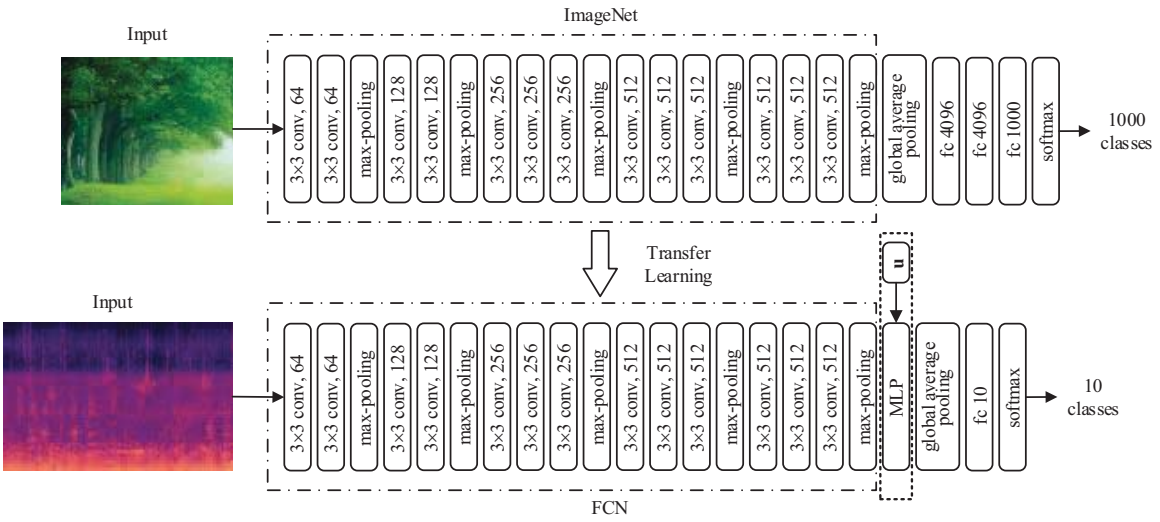


Figure 2: *Illustration of CNN-based transfer learning (VGGNet as the example).*

tween two quite different classification tasks with the inputs of 2D image signals and 1D audio signals. One strong motivation behind this is the spectrograms of the audio signal can be also considered as the 2D images which are potentially have the similar structures to those samples in the image classification task. Specifically, we conduct the transfer learning mechanism by adopting the pre-trained CNNs with different architectures from the ImageNet task to the DCASE2018 ASC subtask A. We compare several typical CNN architectures and the modified VGG network works the best. Furthermore, by leveraging more input channels and training data fragments, the classification accuracy of our proposed system is increased from 59.7% to 77.8% on the evaluation set, in comparison to the officially provided CNN system trained using only audio data. This demonstrates the transfer learning between image and audio tasks is quite effective. The remainder of the paper is organized as follows. In Section 2, we first introduce the proposed system overview used for experiments. In Section 3, we report and analyze experiment results. Finally we summarize our work and present conclusions in Section 4.

2. Our Proposed ASC System

The overall flowchart of our proposed system is illustrated in Figure 1. For the feature extraction, the audio samples are processed to extract MFCC features. Accordingly, for each audio recording, a Mel spectrogram can be extracted. In the training stage, if the audio recording is quite long, we can split the corresponding Mel spectrogram into several Mel spectrogram fragments for the purpose of easy model training and data augmentation. Moreover, if the audio is recorded by a binaural microphone, the average of the left channel and right channel in the time domain is often adopted as the input. But the original data from the left and right channels can be still used as the augmented training data. With the extracted features of all training data, the CNN-based classifier can be learned via the transfer learning mechanism, which will be elaborated in the next section.

In the testing stage, for each channel, we first extract the Mel spectrogram features and split into more fragments. Then, using the classifier, the score of each channel is calculated by averaging the posterior probabilities of all fragments. Finally, the score fusion is performed to generate the detected scenes by averaging the scores of all channels.

3. CNN-based Transfer Learning

CNN has been widely used for acoustic scene classification [20, 21]. Compared with traditional handcrafted feature extraction, CNN can automatically learn the feature representation. And it has been proved that CNN-based system can often achieve a better accuracy compared with the traditional machine learning method on the acoustic scene classification task [19]. The basic components of CNN are convolution, pooling and activation layers. The typical CNNs, including AlexNet [22]. Oxford VGGNet [23], ResNet [24] take fixed-size input. In this study, as shown in Figure 2, we turn the AlexNet/VGGNet/ResNet into a fully convolutional network (FCN) by simply removing its fully connected layers. To adopt these networks for acoustic scene classification, a softmax layer is appended with 10 nodes as 10 scene classes. In the image classification task such as ImageNet [22], AlexNet/VGGNet/ResNet achieved a great success. AlexNet won the champion of ILSVRC (ImageNet Large Scale Visual Recognition Competition) in 2012. VGG was the winner of the ILSVRC competition in 2014. ResNet has also won championships of multiple contests in computer vision area recently. This is the reason that we make a comparison among these three CNN structures in the experiment.

By conducting the transfer learning from image classification to acoustic scene classification, we should also consider the adaptation problem of both inputs and outputs between these two different tasks. Different from image classification task often containing RGB three channels, we only use one channel of Mel spectrograms as the input. We also compare the FCNs with random initialization and the pre-trained FCNs based on ImageNet dataset [22].

Based on the FCN part, we investigate two designs of the linking between the FCN and the final output layer with 10 nodes. Assuming that the FCN output is a 3-dimensional array of size $F \times T \times C$, F and T are the sizes regarding to the frequency and time domains. The first design is using a global

average pooling across both F-axis and T-axis to generate a C-dimensional vector which is fully connected (FC) to the output layer.

For the second design, an additional module is adopted as shown in the dotted box of Figure 2. We first reshape the FCN output as a 2-dimensional array $\boldsymbol{A}$ of size $L \times C$ ($L = F \times T$):

$$\boldsymbol{A} = \{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_L\}, \boldsymbol{a}_i \in \mathbb{R}^C. \quad (1)$$

Then, we add a multilayer-perceptron (MLP) layer with C' output nodes and the $\tanh$ as the non-linear activation function. Moreover, a learnable C'-dimensional vector $\boldsymbol{u}$ is used to perform an element-wise product with the MLP output:

$$\boldsymbol{e}_i = \boldsymbol{u} \circ \tanh(\boldsymbol{W}\boldsymbol{a}_i + \boldsymbol{b}), \quad (2)$$

where $(\boldsymbol{W}, \boldsymbol{b})$ is the parameter set of MLP. $\boldsymbol{E} = \{\boldsymbol{e}_1, \cdots, \boldsymbol{e}_L\}$ as a new representation of $\boldsymbol{A}$ is then fed for the global average pooling across the L-axis. In our experiments, $C = 512$ and $C' = 256$. Compared with the first design, our proposed second design is verified to yield a higher accuracy by using more parameters trained with the audio data. Please note that in the subsequent experiments, the first design is used as default and the comparison between the first and second designs is only conducted in the final system.

4. Experiments and Analysis

4.1. Data set and feature extraction

The experiments use the TUT Urban Acoustic Scenes 2018 dataset. The audio recordings with 48 kHz sampling rate containing 10 different scenes were recorded by electret binaural microphone. The length of each audio recording is 10 seconds. For the feature extraction like in [12], we first perform short-time Fourier transform on each audio. The frame length is 46.4ms while the frame shift is 23.2ms. Then, we calculate the amplitude of each bin and apply 128-bank mel-scale filter banks followed by the logarithm transform. Finally the 128-dimensional Mel spectrogram features are generated. Since the audio recording is dual channel, we can use the original data from the left channel, right channel and the average one of left and right channels.

4.2. Different CNN structure comparison

Table 1: *The classification accuracy comparison of different CNN structures and initializations.*

Network	Random-init	Finetune
AlexNet	59.5%	64.3%
VGGNet-16	61.9%	70.6%
ResNet	61.6%	70.2%

Table 1 summarizes the results of AlexNet, VGGNet16, ResNet with different initializations. The system uses only the average channel, does not use the data augmentation (namely splitting one Mel spectrogram of 10s recording into more fragments) and the additional MLP layer in Figure 2. It is interesting to observe the pre-trained CNNs with ImagetNet data always outperform the CNNs with random initialization weights, yielding a significantly absolute increase of 8.7% accuracy in the pre-trained VGGNet-16 case. Moreover, the VGGNet-16 achieves the best performance compared with the other two CNNs for both random initialization and fine-tuning. The reason why VGGNet-16 is superior might be explained as the tradeoff of increasing the network complexity and alleviating the overfitting problem.

It is well known that 1D audio signal is quite different from the 2D image signal. However, ImageNet dataset contains 1281167 image samples, which are far more than the training samples of our ASC task. We suppose that some natural scene images in the ImageNet dataset may include similar structures to the Mel spectrograms, e.g., dusk, walls, roads, and even wheat fields. Therefore, for the FCN part which is extracting distinctive features, the pre-training using ImageNet data can help FCN to extract more useful information from the input Mel spectrograms.

4.3. Data augmentation and segmentation

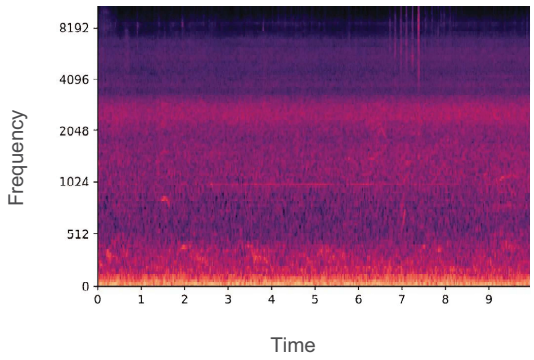
Table 2: *The classification accuracy comparison of different data augmentations using the pre-trained VGGNet-16.*

Network	N_s	Accuracy
VGGNet-16	3	71.6%
VGGNet-16	6	73.9%
VGGNet-16	9	71.8%
VGGNet-16	12	71.8%

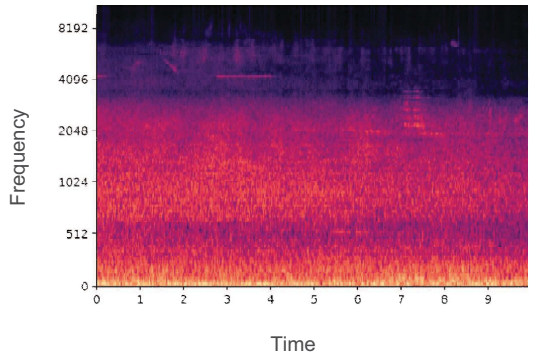
Inspired by [25], we split the whole Mel spectrogram of each 10s audio recording into more fragments. We used the small fragments as the augmented samples for training. And for the testing stage, the score of each audio recording is the average of posterior probabilities of all fragments. Table 2 lists the results of the classification accuracy with different settings of data augmentations based on the pre-trained VGGNet-16 classifier. N_s is the number of fragments for an audio. Compared with the result in Table 1 without splitting, all settings of the number of fragments increased the accuracy. And the setting of 6 fragments achieved the best performance with an absolute accuracy gain of 3.3% over the no-splitting case.

To give a better explanation, we compared and analyzed the results of the different settings of fragments only in the testing stage. We found that the main reason that there was an optimal setting of fragments was the effective segments in a long audio recording which could distinguish from other confusing classes might be a small faction (less than 50%) of the whole recording. In such cases, the no-splitting setting obviously led to the misclassified result. Similarly, if we split one recording into too many fragments (one frame as one fragment in the extreme case), the classification result was still not correct. So there might be a proper setting of the number of fragments where the voting result could be correct.

In Figure 3, two examples of Mel spectrograms labeled as "public_square" and "street_traffic" from the average channel are given as an illustration. These two audio recordings sounds very similar. However, both are classified as the "street_traffic" class. And we found that in Figure 3(b) a horn sound is the most effective segment to distinguish these two classes between 7.2 and 7.6 seconds of this audio. This could be also easily observed in the Mel spectrogram. This further indicates that the decision based a long audio recording should pay more attention to the effective segments, which implies that the detection and extraction of representative segments for acoustic scene classification



(a) The example labeled as public_square.



(b) The example labeled as street_traffic.

Figure 3: *The comparison of Mel spectrograms of two examples labeled as public_square and street_traffic from the average channel.*

are quite important.

4.4. Overall comparison

In this subsection, we further examine the effectiveness of using multi-channel fusion and the newly designed linking between FCN and the output layer as shown in Figure 2. Considering that the audio is recorded by a binaural microphone, both left and right channel could be also used for training as the supplementary of the average channel. In the testing stage, the score fusion is performed for multi-channel inputs.

Table 3: *The classification accuracy comparison of different systems.*

Input	Network	N_s	Accuracy
Average	Baseline [26]	1	59.7%
Average	VGGNet-16	6	73.9%
Left,Right,Average	VGGNet-16	6	75.5%
Left,Right,Average	VGGNet-16+MLP	6	77.8%

Table 3 shows the corresponding experimental results. The multi-channel fusion led to the accuracy increase from 73.9% to 75.5%, demonstrating the complementarity among different channels. Moreover, by using the new design of linking between FCN and output layer, the VGGNet-16+MLP system generated an additional accuracy gain of 2.3% over the VGGNet-16 system. Finally, our best system can increase the accuracy from 59.7% to 77.8% compared with the officially provided baseline system [26]. The baseline system is a convolutional neural network with two layers and random initialization.

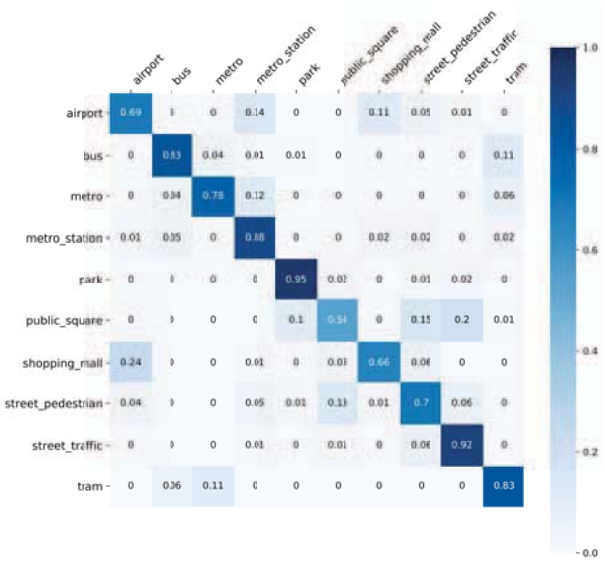


Figure 4: *Confusion matrix of the proposed system . X-axis indicates the predicted label and Y-axis indicates the true label.*

Figure 4 shows confusion matrix of the proposed system. It can be observed that the "public_square" classification results are the worst and are mainly misclassified as "park", "street_pedestrian", and "street_traffic". Also the accuracy of "airport" and "shopping_mall" did not reach 70%. Our future work intends to design more effective segmentation approaches to improve the performance of those confusing scenes.

5. Conclusions

In this paper, we investigated the transfer learning mechanism of CNN architecture and weight initialization from the image classification to acoustic scene classification. We experimented on several typical CNNs (AlexNet, VGGNet-16, ResNet-50) and found that the pre-trained networks from ImageNet task achieved better performance of acoustic scene classification than randomly initialized one. Moreover, by splitting each audio recording into smaller segments, the accuracy could be further improved. Validated on the DCASE2018 ASC subtask A, our proposed system achieved an accuracy of 77.8% on the evaluation set. This demonstrates that the transfer learning between two tasks with quite different input signals (2D image vs. 1D audio) is possible.

6. Acknowledgements

This work was supported in part by the National Key R&D Program of China under contract No. 2017YFB1002202, the National Natural Science Foundation of China under Grants No. 61671422 and U1613211, the Key Science and Technology Project of Anhui Province under Grant No. 17030901005, MOE-Microsoft Key Laboratory of USTC, and Huawei Noah's Ark Lab.

7. References

[1] S. Kornblith, J. Shlens, and Q. V. Le, "Do better imagenet models transfer better?" *arXiv preprint arXiv:1805.08974*, 2018.

[2] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[3] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE press, 2006.

[4] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, 2015.

[5] http://www.cs.tut.fi/sgn/arg/dcase2016/.

[6] http://www.cs.tut.fi/sgn/arg/dcase2016/.

[7] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 321–329, 2006.

[8] R. G. Malkin and A. Waibel, "Classifying user environment for mobile applications using linear autoencoding of ambient audio," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, vol. 5. IEEE, 2005, pp. v–509.

[9] W. Nogueira, G. Roma, and P. Herrera, "Sound scene identification based on mfcc, binaural features and a support vector machine classifier," *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*, 2013.

[10] J. T. Geiger, B. Schuller, and G. Rigoll, "Recognising acoustic scenes with large-scale audio feature extraction and svm," *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*, 2013.

[11] B. Elizalde, H. Lei, G. Friedland, and N. Peters, "An i-vector based approach for audio scene detection," *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*, 2013.

[12] Y. Han, J. Park, and K. Lee, "Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification," *the Detection and Classification of Acoustic Scenes and Events (DCASE)*, pp. 1–5, 2017.

[13] Z. Weiping, Y. Jiantao, X. Xiaotao, L. Xiangtao, and P. Shaohu, "Acoustic scene classification using deep convolutional neural network and multiple spectrograms fusion."

[14] K. Patil and M. Elhilali, "Multiresolution auditory representations for scene classification," *cortex*, vol. 87, no. 1, pp. 516–527, 2002.

[15] A. Mesaros, T. Heittola, and T. Virtanen, "Tut database for acoustic scene classification and sound event detection," in *Signal Processing Conference (EUSIPCO), 2016 24th European*. IEEE, 2016, pp. 1128–1132.

[16] S. Mun, S. Park, D. K. Han, and H. Ko, "Generative adversarial network based acoustic scene training set augmentation and selection using svm hyper-plane," Tech. Rep., DCASE2017 Challenge, Tech. Rep., 2017.

[17] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2392–2396.

[18] H. Eghbal-Zadeh, B. Lehner, M. Dorfer, and G. Widmer, "Cp-jku submissions for dcase-2016: A hybrid approach using binaural i-vectors and deep convolutional neural networks," *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2016.

[19] Y. Han and K. Lee, "Convolutional neural network with multiple-width frequency-delta data augmentation for acoustic scene classification," *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*, 2016.

[20] R. Hyder, S. Ghaffarzadegan, Z. Feng, and T. Hasan, "Buet bosch consortium (b2c) acoustic scene classification systems for dcase 2017 challenge."

[21] J. Xu, Y. Zhao, J. Jiang, Y. Dou, Z. Liu, and K. Chen, "Fusion model based on convolutional neural networks with two features for acoustic scene classification."

[22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[24] 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.

[25] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," *Proc. Interspeech 2017*, pp. 1089–1093, 2017.

[26] https://github.com/DCASE-REPO/dcase2018_baseline/tree/master/task1.