Two-Stage Classification Learning for Open Set Acoustic Scene Classification

Chunxia Ren[1] and Shengchen Li[2(✉)]

[1] Beijing University of Posts and Telecommunications, No. 10 Xitucheng Road, Haidian District, Beijing, China
chunxiaren@bupt.edu.cn

[2] Department of Intelligent Science, School of Advanced Technology, Xi'an Jiaotong-Liverpool University, 111 Ren'ai Road, Suzhou Industrial Park, Suzhou 215123, Jiangsu Province, P. R. China
shengchen.li@xjtlu.edu.cn

Abstract. Most of the research on acoustic scene classification (ASC) focuses on classification problem with only known scene classes. In practice, scene classification problem to be solved generally is based on an open set, which contains unknown scenes. This paper proposes a two-stage method that solves the open set problem on ASC. The proposed system decomposes open set ASC problem into two stages. To mitigate the impact of unknown scenes on the subsequent recognition process of known scenes, the first stage is to identify unknown scenes. The second stage classifies defined acoustic scenes. In this case, the threshold selection strategy we proposed further sorts out unknown scenes that were not identified in the previous stage. Experiments show that the method proposed in this paper can effectively identify unknown scenes and classify known scenes, by segmenting the open set acoustic scene classification task and selecting an appropriate judgment threshold. On the development dataset released by DCASE Challenge 2019 Task 1C, the model proposed outperforms the first place.

Keywords: Acoustic scene classification · Open set · Two-stage classification · Threshold selection strategy

1 Introduction

As an environmental identification problem, acoustic scene classification (ASC) attracts growing attention [2]. ASC processes the audio signal and then extracts feature information, and the scene is identified by event or semantic information contained in feature representation [16]. ASC is widely used in smart wearable devices, robots, home surveillance and security systems, environmental noise monitoring.

The challenge of Detection and Classification of Acoustic Scenes and Events (DCASE) [15] provides a series of the open-source database and evaluation methods which develop ASC. In recent years, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) and Convolutional Recurrent Neural Networks (CRNNs) are recognised as effective models for ASC problems and

X. Shao et al. (Eds.): CSMT 2020, LNEE 761, pp. 124–133, 2021.
https://doi.org/10.1007/978-981-16-1649-5_11

are generally superior to traditional machine learning methods in performance [1,17,22]. However, the research of ASC is mainly focused on the closed set, that is, the scene classes used in the testing phase and the training phase are same. In practical scene analysis application, undefined scene classes other than limited known scene classes are often encountered. Thus, open set recognition task [4–6] which needs to additionally identify the undefined scenes as an unknown class is more useful despite higher complexity.

This paper focuses on solving open set ASC problem. The significant differences in data composition between open set and closed set make traditional ASC models no longer applicable to open set ASC. To the best of our knowledge, the research based on open set classification tasks is mainly based on one-stage classification methods [7,8,18]. Daniele et al. [3] firstly proposed a solution to the open set problem in the ASC field. They not only use Support Vector Data Description (SVDD) classifier to learn a hypersphere from known scenes to distinguish unknown class but also introduce a new protocol and indicator for evaluating the open set ASC task. The introduction of this question has attracted some scholars to study.

The DCASE Challenge 2019 Task 1C further facilitates extensive research in open set ASC. These solutions proposed by Wilkinghoff et al. [20] and Lehner et al. [13] classify known classes and separate unknown classes only by learning known classes in a single classification system; the difference is that the former used Deep Convolutional Auto-Encoders (DCAEs) as classification model and the latter used the improved ResNet variant [11,12] as the classifier. These one-stage classification methods [13,20] in which unknown classes do not participate in training phase pay more attention to the inter-class differences of known scenes but are not necessarily useful for separating unknown scene from known scenes. A one-stage classification method proposed by Zhu et al. [23] is to put the unknown class into training phase and designs an $K+1$ classifier that treats the unknown class like K known classes. This method uses CRNN-Attention mechanism model [19,21] as the classifier and achieves the first place of the DCASE Challenge 2019 Task 1C. There is a problem with this method. Although the unknown classes participate in training process, the operation where unknown classes are unreasonably regarded as a known scene is likely to ignore the difference between the unknown class and the entire set of known classes in the distribution of the feature space.

To avoid the problems of the two types of methods mentioned above [14, 20,23], this paper proposes a solution for open set ASC. Unknown classes are no longer considered to be an equal role for K known classes in this paper. Consequently, this paper designs a two-stage classification learning system for open set ASC to better solve the open set classification problem. The first stage is used to distinguish unknown classes from the entire set of known classes, reducing the impact on the next stage. The second stage is used to further divide the known classes into defined scene labels and separate the remaining unknown classes which not identified during first one, as well as we proposed a threshold selection strategy to assist in identification of unknown classes. Experiments show that our proposed two-stage method which identifies unknown classes and

classifies known classes more precisely than traditional one-stage methods is an effective open set ASC solution.

The remainder of this paper is organized as follows: Sect. 2 describes the two-stage classification method for open set ASC presented in this paper; Sect. 3 introduces experimental setup and results in analysis, and Sect. 4 summarizes current works and discusses future research directions.

2 Proposed Two-Stage Classification Model

There are two problems to be solved in the open set ASC task, one is to identify unknown scene, and the other is to classify known scenes. Therefore, in this section, a two-stage classification learning model for open set ASC is proposed. The system divided the open set ASC question into two parts and resolves them in two stages. This section describes in detail how the two-stage ASC system (as shown in Fig. 1) proposed for the open set completes the classification task.

2.1 Two-Stage Classification Model—The First Stage

The two-stage classification system for open set ASC is proposed in this paper (as shown in Fig. 1). The main role of the first stage is to separate the same or similar unknown class encountered in training phase, these separated unknown class no longer participates in the second stage testing phase, reducing the impact on the second stage classification.

There are two reasons for unknown class to participate in the first stage training phase. One is to make the model not overemphasize discriminative features

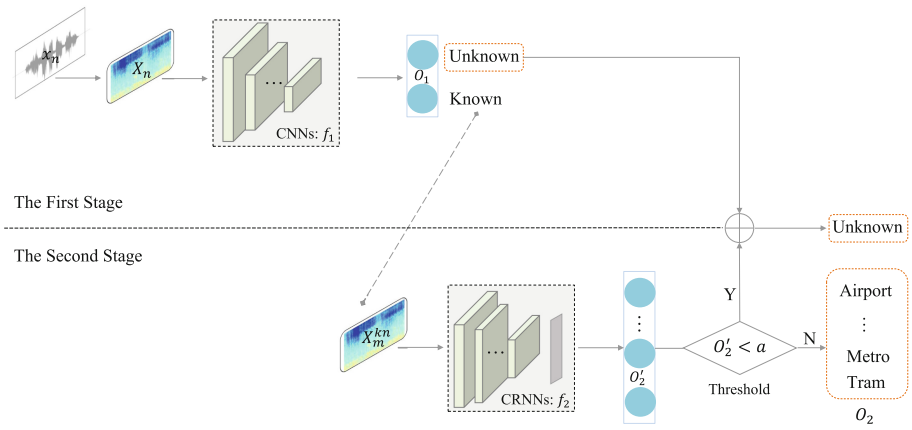

Fig. 1. A two-stage open set ASC system proposed is consist of the first stage and the second stage classification models. The output O_2 is the sum of K known classes obtained by the second stage classification model and unknown class obtained by the first stage and the second stage classification models.

of the known classes by adding the unknown class samples; the other is to fully consider the situation that may occur during testing phase because it does not know in advance whether the sub-scenes contained in the unknown class of the testing phase have already been encountered in the training phase.

To classify the scenes into the known class or unknown class, CNNs which performed well in ASC task is used as the first stage classification model [17,22]. After decomposing the original open set ASC task into two problems solved in two stages respectively, the task complexity of the first stage is significantly reduced, then the requirement for model complexity is also reduced. Thus, the first stage task can be accomplished using a CNN classification model with shallow structure. The CNNs proposed in this paper is composed of four convolution layers followed by maxpooling layer.

In the first stage (Fig. 1 upper part), the features X_n which represented by the log-mel spectrogram of the scene audio signal x_n is taken as input, where n represents the index of audio. The advanced feature representation of the original input X_n is extracted by the shallow CNNs with four layers. The distinctive information of the features of the known class and the unknown class that appeared in the training phase is learned and used as a classification basis. The global average pooling (GAP) is used to convert the feature map of the last layer of CNNs into feature points by averaging pooling. Thus, feature points with significant visibility in CNNs are reserved by the GAP. The output p_n of the neural network is a predicted probability that indicates whether the sample belongs to the unknown class. To this end, the model is optimized by updating the weights during backpropagation and minimizing the binary cross-entropy loss:

$$l = -\sum_{n=1}^{N}((y_n \log p_n) + (1 - y_n)\log(1 - p_n)) \tag{1}$$

where N is the number of samples in training phase, y_n represents the estimated label of the nth sample. Finally, the samples of the first stage are expected to be classified as known class or unknown class.

2.2 Two-Stage Classification Model—The Second Stage

Compared with the first stage, the second stage needs to detailly classify the complex known scenes into K defined classes. From the perspective of task complexity, the second stage is more complicated, which may result in shallow CNNs does not necessarily complete the task well. Since CRNNs was proposed by [1], there has been a lot of work to prove its excellent performance on ASC. Therefore, we use it as the classification model in the second stage [21].

As shown in the lower part of Fig. 1, the features $X_m^k n$ of the known scenes $x_m^k n$ is used as input to CRNNs, where m is denoted as the index of audio during the second stage. Among CRNNs, CNNs which acts as advanced features extractor passed the abstracted advanced feature information into bi-directional RNN (Bi-RNN). The information in features that helps to classify scenes is not only independent, but it is also sometimes related to its occurrence time. Therefore,

considering the need to maintain the temporal resolution of the sequence generated by Bi-RNN, the pooling operation only occurs on the frequency axis. In this way, Bi-RNN learned the contextual timing relationship of the feature and encode it. Bi-RNN regarded as another advanced feature extractor, but different from CNNs mode.

In the first stage, some unknown scenes used for the testing phase that differs greatly from the trained unknown scenes in the feature space may be missed. Since the classifier in the second stage is designed for known scenes, it could be assumed that the probability of unknown scenes is relatively low. Therefore, a judgment threshold h is needed to determine the scene with an output probability below h as an unknown class. To make the threshold at this stage divide the known and unknown classes more scientifically, we propose a threshold selection strategy. If M_{uk} which is the number of unknown samples in the testing phase is known, the choice of threshold hs should make predicted probability of at least M_{uk} testing samples lower than h. When M_{uk} is unknown, but the relationship between the accuracy of the unknown classes and the accuracy of the system is known as:

$$ACC = (1 - \beta) * ACC_{kn} + \beta * ACC_{uk} \tag{2}$$

Then the value of the threshold should result in the predicted probability of $\beta * N_t$ samples being lower than h, where β is a weight coefficient which less than 1 but over 0 and N_t is the number of testing samples.

The weighted average operation proposed by [21] is used in the second stage to obtain a suitable probability output so that the selected threshold h separates the unknown class which is not identified in the first stage. The weighted average operation is following,

$$O_2' = \frac{\sum_{t=0}^{T-1} O'(t)}{\sum_{t=0}^{T-1} Z_{soft}(t)} \tag{3}$$

where

$$O'(t) = Z_{soft}(t) \odot Z_{sigm}(t) \tag{4}$$

T is the frame-level resolution and O' is the element-wise multiplication of the outputs of two fully connected layers whose activation function is softmax Z_{soft} and sigmoid Z_{sigm}.

Then, the output corresponding to the first stage is $O_1\{O_1^{kn}; O_1^{uk}\}$, the output of the second stage is O_2', and the further output after the threshold a judgment is $O_2\{O_2^{kn}; O_2^{uk}\}$. The output O composition of our proposed system should be the combination of the sum of the unknown class identified in the first stage and the second stage, and the classification results of the K known classes in the second stage as following,

$$Output\colon O = (O_2^{kn}; O_1^{uk} \bigcup O_2^{uk}) \tag{5}$$

3 Experiments

3.1 Dataset and Experimental Setup

This paper verified the two-stage classification system for open set ASC presented on the development dataset published by the Task 1C of DCASE 2019 Challenge. The dataset contains known scenes and unknown scenes; the former is 10 scenes recorded in 10 different European cities, each recording approximately 1440 audio samples; the latter consists of 4 different sub-scenes, the number of audio samples recorded in each scene is about 480. The duration of the audio samples is 10 s.

The ratio of the training set and the testing set is 3:1. 10 visible sub-scenes of known class appear in both training and testing sets. There are two possible situations where invisible sub-scenes of unknown class in the testing set may be completely different from sub-scenes of unknown class in the training set or maybe partial duplication. To demonstrate the effectiveness of the proposed system, we do a set of comparative experiments. The parameter setups in the experiment are as follows.

Figure 2 shows the composition of the classification models in the first and second stages. Log-mel spectrogram is used as the features of audio samples, with 640 frames per chunk by 128 mel bins, and then each chunk is evenly divided into 5 segments, each segment has 128 frames. Batch normalization [9] is applied after each convolutional layer. During the experiment, dropout was added to avoid over-fitting of the proposed model, the judgment threshold was chosen to be 0.2 by threshold selection strategy, and the Adam optimizer with learning rate which fixed at 0.001 is used.

3.2 Results and Analysis

The number of correctly classified audio samples in the total number of audio samples called classification accuracy is used as the score of the open set ASC. Accuracy is calculated as the weighted average of the known classes and unknown class, as shown below:

$$ACC_{weighted} = 0.5 * ACC_{kn} + 0.5 * ACC_{uk} \tag{6}$$

where known classes accuracy ACC_{kn} is the average of the class-wise accuracy.

In Table 1, this experiment compares the proposed model with two typical one-stage models on the development dataset divided by the Task 1C of DCASE 2019 Challenge. These two typical models are the Baseline and the best model [23] published on DCASE 2019 Challenge, respectively. Among them, the 10 classification model based on CNNs is adopted by the Baseline, with 0.5 as the judgment threshold; the 11 classifications based on CRNN-Attention model [19] is adopted by the model [23]. And our proposed two-stage classification learning system achieves better results with nearly 5% improvement over the best model by identifying unknown classes in the first stage, classifying known classes

First stage （CNNs）	Second stage （CRNNs）
Log-mel spectrogram 128 frames * 128 mel bins	Log-mel spectrogram 128 frames * 128 mel bins
$\left[\binom{3*3\ @64}{BN, ReLU}, P(1,2)\right] * 2$	$\left[\binom{3*3\ @128}{BN, ReLU}, P(1,2)\right] * 5$
$\left[\binom{3*3\ @128}{BN, ReLU}, P(1,2)\right] * 2$	$\left[\binom{3*3\ @64}{BN, ReLU}, P(1,4)\right]$
Global average pooling	Bi-GRU @128
	Weighted average

Fig. 2. The models of the first and second stages. 'P' represents pooling, 'BN' is Batch Normalization.

unknown classes that are difficult in the previous stage in the second stage. Both the average accuracy of known classes and the accuracy of the unknown class are higher than the one-stage classification methods, which proves that our proposed two-stage method is more reasonable and has better performance for unknown classes recognition and known classes classification. To further compare the difference in the class-wise accuracy between the system proposed in this paper and the other two systems, Table 1 shows the comparison of the class-wise average accuracy of scene classes on these three models. It can be seen that the proposed system has the highest accuracy in multiple scene classes, but it does not perform well in individual classes such as "Airport", "Street_pedestrian", and "Tram". One possible reason is that these low-accuracy scenes are similar to other scenes in the feature space, causing the system to misjudge.

Compared with several other models that use a fixed empirical threshold, the model we proposed verifies the rationality of the threshold selection strategy. The model [10] and Baseline in Table 2 utilize a traditional threshold of 0.5 to identify samples with prediction probability lower than 0.5 as "Unknown". This choice leads to the randomness of results, and it is difficult to ensure that 0.5 is the appropriate probability boundary between the known classes and the unknown classes. These models do not take into account the probability of prediction and data composition together. The threshold of model [23] is selected as 0.4, and the same problem exists. We choose the threshold as 0.2 based on the threshold selection mechanism proposed in this paper.

Since the number of samples of the unknown class in the testing phase is 345, accounting for nearly 7.6% of the testing sample. 0.2 is selected as the threshold according to the threshold selection strategy so that the prediction

Table 1. The accuracy (%) of the corresponding model. Among them, "Known" represents the average accuracy of 10 known scene classes, "Unknown" represents the accuracy of unknown scene classes, and "Overall" is the accuracy calculated by Formula (6).

Accuracy	Model		
	Baseline	Zhu et al. [23]	Our model
Airport	44.2	65.3	41.1
Shopping_mall	50.9	26.3	**71.7**
Metro_station	41.3	42.1	**56.6**
Public_square	34.7	39.8	**45.0**
Metro	51.5	42.3	**51.7**
Tram	60.7	57.6	55.1
Street_pedestrian	47.5	37.3	46.9
Street_traffic	78.4	74.4	**80.4**
Bus	59.3	52.3	53.7
Park	74	80.8	64.8
Known	54.3	51.8	**56.7**
Unknown	43.1	75.9	**80.3**
Overall	48.7	63.9	**68.5**

Table 2. The relationship between selected threshold and accuracy (%) of the corresponding model. "Unknown" represents the accuracy of unknown scene classes, and "Overall" represents the accuracy calculated by Formula (6).

Model	Threshold	Unknown	Overall
Baseline	0.5	43.1	48.7
Kong et al. [10]	0.5	48.1	53.1
Zhu et al. [23]	0.4	75.9	63.9
Our model	**0.2**	**80.3**	**68.5**

probability of about 7.5%–8% testing samples is lower than 0.2. As is seen from the above Table 2, the method we proposed has the highest accuracy of the unknown class: 80.3%, which has an obvious advantage than other methods. The above scheme is obtained when the number of "Unknown" samples is known. We verified the rationality of the proposed threshold selection mechanism on the dataset of private Kaggle leaderboard when the number of "Unknown" samples is unknown. Therefore, according to the formula (6), we choose a threshold of 0.4, which makes about half of the testing samples' prediction probability is lower than the threshold. Under the threshold selection strategy, we achieved the best results of the private Kaggle leaderboard.

4 Conclusion

This paper proposes a two-stage classification learning solution for open set ASC, which achieves 68.5% by using the proposed model on the development dataset of the DCASE 2019 for open set ASC, which is better than the optimal performance released by DCASE Challenge 2019 Task 1C. The experiment proves that the proposed model is really useful. In the future, we will explore how to improve the accuracy of the known classes while ensuring the unknown class accuracy, and balance the accuracy of known classes and unknown class. Besides, we will also study the less dependent experience-based solutions for open set ASC, and the feature representation methods that can more clearly distinguish different scenes.

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China (62001038).

References

1. Adavanne, S., Politis, A., Nikunen, J., Virtanen, T.: Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. IEEE J. Sel. Top. Signal Process. **13**(1), 34–48 (2018)
2. Barchiesi, D., Giannoulis, D., Stowell, D., Plumbley, M.D.: Acoustic scene classification: classifying environments from the sounds they produce. IEEE Signal Process. Mag. **32**(3), 16–34 (2015)
3. Battaglino, D., Lepauloux, L., Evans, N.: The open-set problem in acoustic scene classification. In: 2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC), pp. 1–5. IEEE (2016)
4. Bendale, A., Boult, T.: Towards open world recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1893–1902 (2015)
5. Bendale, A., Boult, T.E.: Towards open set deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1563–1572 (2016)
6. Chen, J., Sathe, S., Aggarwal, C., Turaga, D.: Outlier detection with autoencoder ensembles. In: Proceedings of the 2017 SIAM International Conference on Data Mining, pp. 90–98. SIAM (2017)
7. Chen, Y., Zhou, X.S., Huang, T.S.: One-class SVM for learning in image retrieval. In: ICIP, vol. 1, pp. 34–37. Citeseer (2001)
8. Ding, Z., Fei, M.: An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window. IFAC Proc. Vol. **46**(20), 12–17 (2013)
9. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
10. Kong, Q., Cao, Y., Iqbal, T., Xu, Y., Wang, W., Plumbley, M.D.: Cross-task learning for audio tagging, sound event detection and spatial localization: DCASE 2019 baseline systems. arXiv preprint arXiv:1904.03476 (2019)

11. Koutini, K., Eghbal-Zadeh, H., Dorfer, M., Widmer, G.: The receptive field as a regularizer in deep convolutional neural networks for acoustic scene classification. In: 2019 27th European signal processing conference (EUSIPCO), pp. 1–5. IEEE (2019)
12. Koutini, K., Eghbal-zadeh, H., Widmer, G., Kepler, J.: CP-JKU submissions to DCASE 2019: acoustic scene classification and audio tagging with receptive-field-regularized CNNs. In: Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), New York, NY, USA, pp. 25–26 (2019)
13. Lehner, B., Koutini, K., Schwarzlmüller, C., Gallien, T., Widmer, G.: Acoustic scene classification with reject option based on resnets. In: Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), New York, NY, USA, pp. 25–26 (2019)
14. Lei, C., Wang, Z.: Multi-scale recalibrated features fusion for acoustic scene classification (2019)
15. Mesaros, A., Heittola, T., Benetos, E., Foster, P., Lagrange, M., Virtanen, T., Plumbley, M.D.: Detection and classification of acoustic scenes and events: outcome of the dcase 2016 challenge. IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP) **26**(2), 379–393 (2018)
16. Phaye, S.S.R., Benetos, E., Wang, Y.: Subspectralnet–using sub-spectrogram based convolutional neural networks for acoustic scene classification. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 825–829. IEEE (2019)
17. Ren, Z., Kong, Q., Han, J., Plumbley, M.D., Schuller, B.W.: Attention-based atrous convolutional neural networks: Visualisation and understanding perspectives of acoustic scenes. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 56–60. IEEE (2019)
18. Scheirer, W.J., Jain, L.P., Boult, T.E.: Probability models for open set recognition. IEEE Trans. Pattern Anal. Mach. Intell. **36**(11), 2317–2324 (2014)
19. Wang, J., Li, S.: Self-attention mechanism based system for dcase2018 challenge task1 and task4. In: IEEE AASP Challenge on DCASE 2018 Technical Reports (2018)
20. Wilkinghoff, K., Kurth, F.: Open-set acoustic scene classification with deep convolutional autoencoders. In: Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE) (2019)
21. Xu, Y., Kong, Q., Wang, W., Plumbley, M.D.: Large-scale weakly supervised audio classification using gated convolutional neural network. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 121–125. IEEE (2018)
22. Yang, Y., Zhang, H., Tu, W., Ai, H., Cai, L., Hu, R., Xiang, F.: Kullback–Leibler divergence frequency warping scale for acoustic scene classification using convolutional neural network. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 840–844. IEEE (2019)
23. Zhu, H., Ren, C., Wang, J., Li, S., Wang, L., Yang, L.: DCASE 2019 challenge task1 technical report. Technical report, DCASE 2019 Challenge, Technical report (2019)