DISTILLING THE KNOWLEDGE OF SPECIALIST DEEP NEURAL NETWORKS IN ACOUSTIC SCENE CLASSIFICATION

Jee-weon Jung, Hee-Soo Heo*, Hye-jin Shim, and Ha-Jin Yu*†

School of Computer Science, University of Seoul, South Korea

ABSTRACT

Different acoustic scenes that share common properties are one of the main obstacles that hinder successful acoustic scene classification. Top two most confusing pairs of acoustic scenes, 'airport-shopping_mall' and 'metro-tram' have occupied more than half of the total misclassified audio segments, demonstrating the need for consideration of these pairs. In this study, we exploited two specialist models in addition to a baseline model and applied the knowledge distillation framework from those three models into a single deep neural network. A specialist model refers to a model that concentrates on discriminating a pair of two similar scenes. We hypothesized that knowledge distillation from multiple specialist models and a pre-trained baseline model into a single model could gather the superiority of each specialist model and achieve similar effect to an ensemble of these models. In the results of the Detection and Classification of Acoustic Scenes and Events 2019 challenge, the distilled single model showed a classification accuracy of 81.2 %, equivalent to the performance of an ensemble of the baseline and two specialist models.

Index Terms— Acoustic scene classification, Specialist models, Knowledge distillation, Teacher-student learning, Deep neural networks

1. INTRODUCTION

Recently, various studies on acoustic scene classification (ASC) systems have been being conducted upon increasing demand from several different industries. The Detection and Classification of Acoustic Scenes and Events (DCASE) challenge is providing a common platform for various studies to compare and examine proposed methods [1, 2]. Based on the efforts of the organizers of the challenge, many different types of research have been conducted to improve the performances of ASC systems. In [3], a sophisticated training procedure for an ASC system was proposed. Other studies have focused mainly on investigating feature extraction and data augmentation techniques for ASC tasks [4, 5]. With such studies and the annual DCASE challenge, the performance of ASC systems has incrementally increased each year. However, to our knowledge, there have been few studies that have analyzed errors that occur due to the characteristics of the ASC task. We believe that such an analysis of the task errors is necessary in addition to designing an elaborate system.

In ASC tasks, common acoustic properties among the different acoustic scenes are a known obstacle that degrade the performance of developed systems [6]. These acoustic properties evoke a few frequently misclassified pairs of acoustic scenes. For more details, Figure 1 (a) shows the proportion of each class among the total misclassified audio segments that use the baseline ASC system. In this figure, we verified that the three most difficult classes to identify occupy more than half of the total error. In addition, Figure 1 (b) shows that most of the errors from the frequently misclassified classes are due to a certain confusing pair. For example, most of the errors from the two classes 'public_square' and 'street_pedestrian,' which were frequently misclassified classes, were caused by the

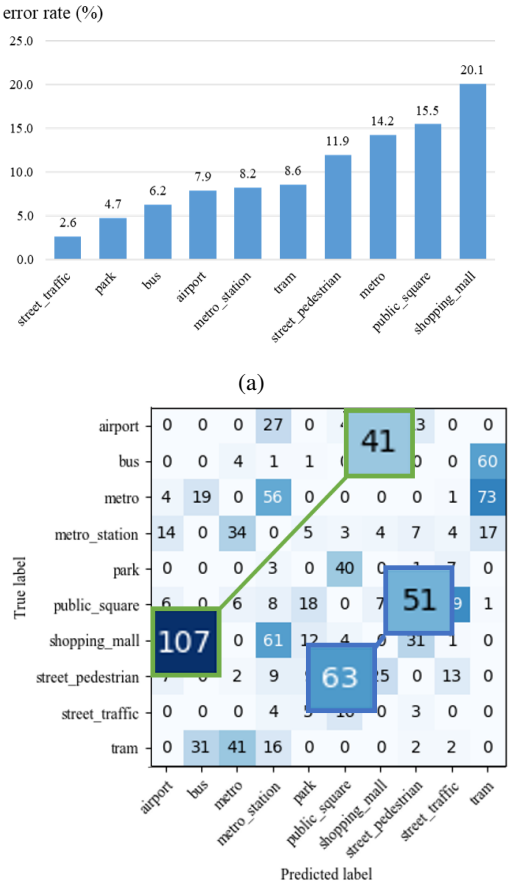

Figure 1: (a) The proportion of each class among the total errors of the baseline model (b) Illustration of the confusion matrix from the baseline model.

*Equal contribution.

† Corresponding author.

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Planning(2017R1A2B4011609)

https://doi.org/10.33682/gqpj-ac63

confusion between each other.

The phenomenon of a few misclassified acoustic scenes severely degrading the overall performance has been alleviated by adopting a knowledge distillation scheme with a *'soft-label'* that models the common properties of acoustic scenes [6]. In this study, we further alleviated this problematic phenomenon by using a *'specialist'* scheme. A specialist model refers to a model that concentrates more on a specific pair of frequently misclassified classes. However, when adopting specialist models, there are still some issues, such as the growing number of parameters (model capacity) and the number of required specialists. To overcome these issues, we further utilized the knowledge distillation scheme combined with the specialist models.

The scheme used in this study distilled the knowledge from the baseline model and two specialist models into a student model. In this scheme, the number of parameters in the distilled model was identical to that in the baseline model. Experimental results on the DCASE 2019 task 1 competition demonstrated that one distilled models shows a performance equal to that of the ensemble of all other models. The main contributions of this paper can be summarized as follows:

1. Adoption of specialist models for frequently misclassified pairs of acoustic scenes.
2. Application of knowledge distillation from a baseline model and two specialist models into one single distilled model in acoustic scene classification.

The rest of this paper is organized as follows: the knowledge distillation (also referred to as teacherstudent learning) framework is introduced in Section 2. Section 3 describes the specialist models and how it is used in this study. The experimental settings and results are detailed in Sections 4 and 5, respectively, and the paper is concluded in Section 6.

2. KNOWLEDGE DISTILLATION IN THE ASC TASK

Knowledge distillation (KD) is a framework where the '*soft-label*' extracted from a DNN is used to train the other DNN (this framework is also referred to as the teacher-student framework) [7, 8]. We refer to the DNN that provides the soft-label as the teacher DNN, and the DNN that is trained using the soft-label is referred to as the student DNN for clarity throughout this paper.

The KD framework was conducted with the following steps. First, a teacher DNN was trained using the categorical cross-entropy (CCE) objective function. After training of the teacher DNN was complete, its parameters were frozen, and only used for providing soft-labels, which were used to train the student DNN. Note that we initialized the student DNN using the parameters from the teacher DNN. The KD framework has been successfully applied to many tasks [9, 10]. It is important to design the teacher DNN to be superior by considering the work flow of the KD framework that trains the student DNN using the output of the teacher DNN. For example, a larger capacity for a model compression task [8], or close talk utterance input for far-field compensation [10] make the teacher DNN superior.

In the ASC task, Heo *et al.* [6] first adopted the KD framework to model the common properties among different acoustic scenes using soft-labels. For example, babbling sounds that occur in both shopping_mall and airport (pre-defined acoustic scenes of the DCASE 2019 challenge) are sometimes labeled as shopping_mall but labeled as airport at other times using a hard-label scheme. Using the KD framework, soft-labels were hypothesized to model these correlations between pre-defined labels based on their common acoustic properties. This approach was successful in that not only was the overall classification accuracy increased but also the number of misclassified audio segments in the most frequently misclassified pair of scenes significantly decreased.

3. KNOWLEDGE DISTILLATION WITH SPECIALIST MODELS

3.1. Specialist Knowledge Distillation

In the KD framework that involves specialist models [7], soft-labels extracted from multiple teacher DNNs were exploited to train a student DNN. In this framework, multiple teacher DNNs comprise one baseline model and a defined number of specialist models. Here, the specialist model refers to a DNN that classifies a subset of classes assigned by a clustering algorithm (e.g. in [7], 300 detailed classes that are in the bird category were set to a specialist model among a total of 15000 categories from *Google*'s internal dataset).

The training process of specialist knowledge distillation is as follows. First, we train the baseline model using a CCE objective function. Next, a defined number of specialist models are initialized using the weight parameters of the baseline model (DNN architecture is identical except for the output layer). Each specialist model is then trained using the CCE objective function with defined subset labels. Finally, the student DNN is trained using multiple soft-labels each extracted from the baseline and specialist models. The loss function $\mathcal{L}_{KD}$ for the training of the student DNN model can be defined as follows:

$$\begin{aligned}&\mathcal{L}_{KD}(\theta;\theta_b,S)\\&\quad=-\sum_{i=1}^{N}\sum_{j=1}^{M} logQ(j|\boldsymbol{x}_i;\theta)[Q(j|\boldsymbol{x}_i;\theta_b)+\sum_{\theta_s\in S}Q(j|\boldsymbol{x}_i;\theta_s)],\end{aligned} \tag{1}$$

$$Q(j|\boldsymbol{x};\theta)=\frac{exp(z_j/T)}{\sum_i exp(z_i/T)}, \tag{2}$$

where N and M denote the size of the mini-batch and the acoustic scenes in the training set, respectively, each input audio segment is referred to as $\boldsymbol{x}_i$, $Q(j|\boldsymbol{x}_i,\theta)$ denotes the posterior probability for the $j'th$ acoustic scene using the concept of temperature T [7], θ_b is the set of parameters in the baseline model, θ_s is the parameter set of specialist model s, and S is the set of specialist models. The function $Q(j|\boldsymbol{x}_i,\theta)$, defined in Eq. (2), has the role of smoothing the results of applying the softmax function to the output $\boldsymbol{z}$ of the DNN. A loss function, $\mathcal{L}_{KD}$, has been proposed to train the single model that can achieve the same as the ensemble of models that have different characteristics [7, 8].

3.2. Specialist Knowledge Distillation for the ASC Task

In this sub-section, we introduce the modifications we make on the knowledge distillation framework with specialists to suit the ASC task. First, we fixed the number of classes to classify rather than selecting a subset of classes. In Hinton *et al.* [7], the number of total classes was too large, making selection of a subset of classes necessary. However, the DCASE 2019 challenge dataset defines ten classes.

Second, in our configuration, one specialist model concentrated on classifying one pair of frequently misclassified acoustic scenes.

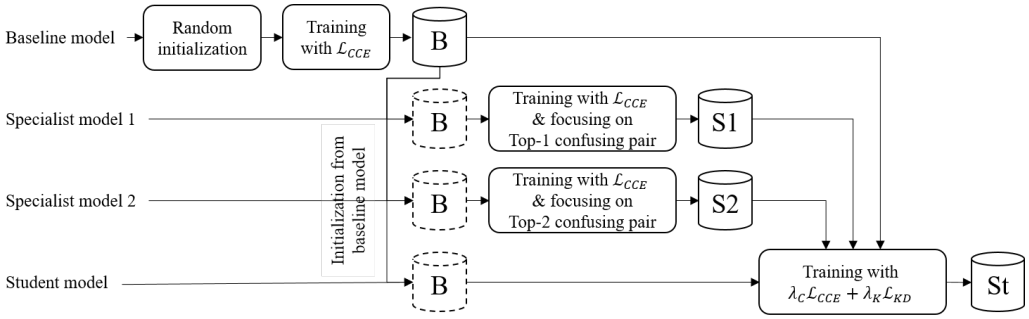

Figure 2: Workflow of the training procedure.

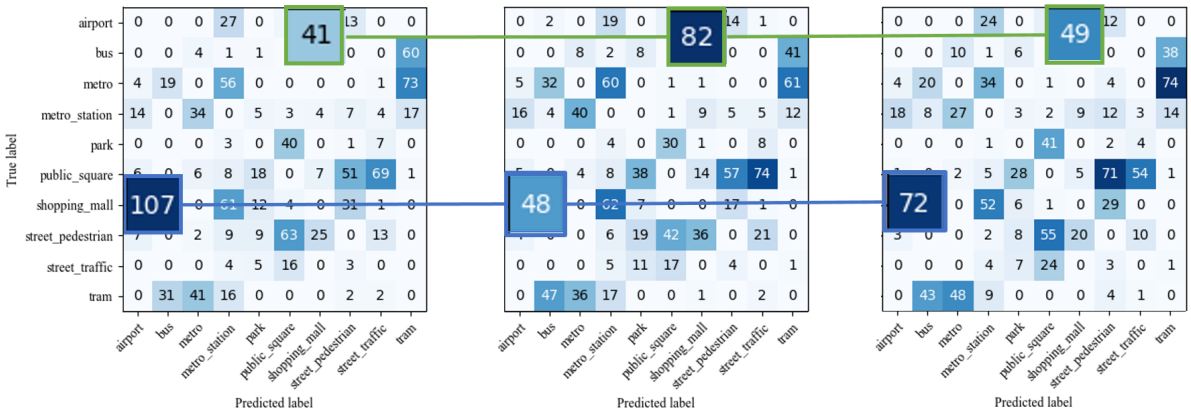

Figure 3: Illustration of the performance analysis based on confusion matrices (right: baseline model, middle: first specialist model, right: student model).

We use two specialist models, where the pair of acoustic scenes to concentrate on is decided based on the confusion matrix of the baseline model; the top two most confusing pairs of acoustic scenes are dealt with two specialist models, respectively. This configuration is based on the analysis that few frequently misclassified acoustic scenes occupy the majority of misclassified samples (see Figure 1). To train the specialist model, we construct half of the mini-batch with target pairs to concentrate on, and the other half with pairs of randomly selected samples from other classes.

After training the specialist models, we train the student model using an objective function composed of the function defined in Eq. (1) and the CCE function, as follows:

$$\mathcal{L} = \lambda_C \mathcal{L}_{CCE} + \lambda_K \mathcal{L}_{KD}, \tag{3}$$

where λ_C and λ_K are the weights of $\mathcal{L}_{CCE}$ and $\mathcal{L}_{KD}$, respectively. The CCE function defined by the true label is used to correct errors that may occur in the teacher models. The values of the two weight coefficients were fixed based on the validation results on the DCASE2019 fold-1 configuration.

The overall training process for the framework used in our study is illustrated in Figure 2.

By applying knowledge distillation using the specialist models, we expect two results. First, class-wise accuracy of the top misclassified acoustic scenes should decrease. Second, the superiority of each specialist model regarding a target pair of acoustic scenes should be well distilled into a single student DNN. To observe whether this objective is successfully achieved, we analyze not only the overall accuracy but also the class-wise accuracies and the number of misclassified samples between each pair of target acoustic scenes that the specialist focused on.

4. EXPERIMENTAL SETTINGS

We conducted all experiments using PyTorch, a deep learning library written in Python [11][1].

We used the Detection and Classification of Acoustic Scenes and Events Challenge Task 1-a dataset for all our experiments. This dataset comprises audio segments that were 10 s and were recorded at 48 kHz with 24-bit resolution; each stereo segment was labelled as one of the pre-defined ten acoustic scenes. The dataset was divided into a development and an evaluation set, where the development set comprised 14400 labelled audio segments, and the evaluation set was not revealed. We constructed a four-fold cross-validation setup using all the data and independently trained four systems. The first fold followed the configuration from the DCASE2019 challenge organizer, and the remaining folds were constructed, taking into account the city where each audio segment was recorded.

We built two separate DNNs for each configuration, where one input raw waveforms and the other input log Mel-energy features. In particular, the model for the raw waveform inputs was con-

[1]Codes used for experiments are available at `https://github.com/Jungjee/dcase2019specialistkd`

structed following the ResNet architecture based on 1-D convolutional layers, and the model for Mel-energy was constructed following the ResNet architecture based on 2-D convolutional layers [12, 13]. We exploited a score-level ensemble in which one uses a CNN that inputs raw waveforms and the other uses a CNN that inputs log Mel-energy features. For data augmentation, we applied a mix-up technique [14] defined as

$$\hat{\boldsymbol{x}} = \lambda \boldsymbol{x}_i + (1-\lambda)\boldsymbol{x}_j, \quad (4)$$

$$\hat{\boldsymbol{y}} = \lambda \boldsymbol{y}_i + (1-\lambda)\boldsymbol{y}_j, \quad (5)$$

$$\lambda = B(\alpha, \alpha), \quad (6)$$

where the pair of $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ represent a set of randomly selected input utterances and the corresponding label, respectively, and $B(\alpha, \alpha)$ is the beta distribution with coefficient α [14]. Label $\boldsymbol{y}_i$ is defined by the true label when training with the CCE function and is referred to the output of the teacher DNN when applying the KD framework.

Refer to the authors' technical report [15] for other details regarding the input features, model architectures, and training procedures[2].

5. RESULTS ANALYSIS

Figure 3 depicts the change of mis-classified samples regarding the most confusing pair ('shopping_mall' and 'airport') in three confusion matrices of the baseline, first specialist, and the student (distilled) model. The number of mis-classified samples of the most confusing pair in these three models was 148, 130, and 121 respectively. Comparing the baseline and the specialist, this result first demonstrates that the mis-classified number of audio segments in target confusing pair decrease in the specialist model than the baseline. However, the overall classification accuracy was similar (for Mel-energy, baseline was 74.33 % and the specialist1 was 74.12 %). Comparing the specialist and the student model, the result of misclassified samples from 130 to 121 shows that not only the overall accuracy increases, but the knowledge of the specialist model is successfully distilled.

To verify whether the superiority of each specialist model was actually distilled to the student DNN, we analyzed the accuracy of the overall and top two most frequently confusing pairs of acoustic scenes. Figure 4 shows the results. Note that these results were from the fold-1 and Mel-energy configuration. We found that the overall accuracy of the student DNN was actually higher than those of all other models. Additionally, we confirmed that for each specialist model the class-wise accuracy of the concentrated pairs increased while the accuracies of other pairs decreased, resulting in similar overall accuracy. The class-wise accuracy of the most confusing pairs in the student model is equal to or higher to those that were the focus of each specialist model. According to this result, we concluded that the designed superiority of each specialist was well distilled to the student DNN. The additional results on the fold-1 configuration are demonstrated in Table 1.

The success of the knowledge distillation is further addressed in Table 2, which reports overall classification accuracies on the evaluation set. This table shows the performances on the evaluation set according to the score-level ensemble methods. The ensemble of 'B+S1+S2+St' means combining the outputs from 32 models (four kinds of models × two fold configurations × two types of input features), and the ensemble of 'St' refers to combining the outputs from eight models. The performance of the student DNN was the same as that of the ensemble of the baseline and two specialist models. This result also verifies that the student DNN trained with specialist knowledge distillation better conducted the ASC task with less number of parameters.

[2] http://dcase.community/documents/challenge2019/technical_reports/DCASE2019_Jung_98.pdf

Table 1: Performances of various systems with the fold-1 configuration according to their accuracies (%) (B: baseline model, S1: $1'th$ specialist model, S2: $2'nd$ specialist model, St: student model).

System	B	S1	S2	St
Raw waveform	73.71	74.89	74.53	75.81
Mel-energy	74.33	74.12	74.48	76.15

Table 2: Performances of various systems with the evaluation configuration according to their accuracies (%) (B: baseline model, S1: $1'th$ specialist model, S2: $2'nd$ specialist model, St: student model).

Systems	B+S1+S2+St	St
Accuracy (%)	81.2	81.2

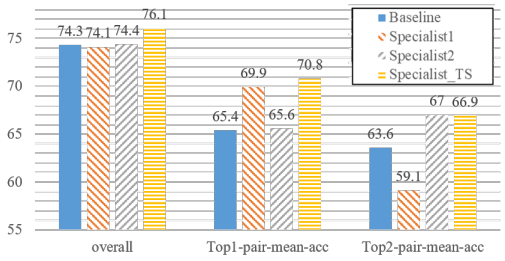

Figure 4: Illustration of the performance analysis based on the mean accuracy of the two classes from the most confusing pairs.

6. CONCLUSION

In this study, we observed that a few pairs of frequently misclassified acoustic scenes occupy more than half of the total misclassified audio segments in an ASC task. For addressing the issue, we adopted the concept of the specialist model, which was designed to concentrate on specific subsets of a task. We modified and trained the specialist models to suit the ASC task. The results show that the specialist model could have not only the superiority that reduces errors for certain confusing pairs but also the inferiority that decreases the discriminative power for other classes. We hypothesized that the KD framework could achieve the identical effect with the ensemble of multiple models by combining superiority into a single model, excluding the inferiority of individual models. The experimental results demonstrated that the KD framework was successful, coherent to our hypothesis and it resulted in overall performance improvements for the ASC system.

7. REFERENCES

[1] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018, pp. 9–13. [Online]. Available: https://arxiv.org/abs/1807.09840

[2] A. Mesaros, A. Diment, B. Elizalde, T. Heittola, E. Vincent, B. Raj, and T. Virtanen, "Sound event detection in the DCASE 2017 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2019, in press.

[3] M. Dorfer and G. Widmer, "Training general-purpose audio tagging networks with noisy labels and iterative self-verification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018, pp. 178–182.

[4] H. Zeinali, L. Burget, and J. H. Cernocky, "Convolutional neural networks and x-vector embedding for DCASE2018 acoustic scene classification challenge," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018, pp. 202–206.

[5] J.-W. Jung, H.-S. Heo, I.-H. Yang, S.-H. Yoon, H.-J. Shim, and H.-J. Yu, "DNN-based audio scene classification for DCASE2017: Dual input features, balancing cost, and stochastic data duplication," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, pp. 59–63.

[6] H.-S. Heo, J.-w. Jung, H.-j. Shim, and H.-J. Yu, "Acoustic scene classification using teacher-student learning with soft-labels," in *INTERSPEECH*, 2019.

[7] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[8] J. Li, R. Zhao, J.-T. Huang, and Y. Gong, "Learning small-size dnn with output-distribution-based criteria," in *Fifteenth annual conference of the international speech communication association*, 2014.

[9] J.-w. Jung, H.-s. Heo, H.-j. Shim, and H.-j. Yu, "Short utterance compensation in speaker verification via cosine-based teacher-student learning of speaker embeddings," *arXiv preprint arXiv:1810.10884*, 2018.

[10] J. Li, R. Zhao, Z. Chen, C. Liu, X. Xiao, G. Ye, and Y. Gong, "Developing far-field speaker system via teacher-student learning," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5699–5703.

[11] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," *NIPS-W*, 2017.

[12] 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.

[13] ——, "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016, pp. 630–645.

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

[15] J.-w. Jung, H.-S. Heo, H.-j. Shim, and H.-J. Yu, "Knowledge distillation with specialist models in acoustic scene classification," DCASE2019 Challenge, Tech. Rep., June 2019.