LOW-COMPLEXITY ACOUSTIC SCENE CLASSIFICATION USING KNOWLEDGE DISTILLATION AND MULTIPLE CLASSIFIERS

Technical Report

Na Zhao

Maxvision, Wuhan, China

ABSTRACT

This technical report describes our submission for Task1a of DCASE2021 challenge. Based on the small-size Mobnet[1] of Tencent team in Dcase2020 task1b, we build our baseline model with only one frequency branch and two classifiers. The two classifers are ten-class classifier and three-class classifier respectively, and they jointly optimize the baseline model. Due to the limitation of model size, we first train a high-accuracy large-size model, and then use distillation method to transfer the knowledge from the large-size model to our baseline model. The final system is quantified from 32-bit float-point to 16-bit float-point.We achieved an accuracy of 59.9% with a model size smaller than 128KB.

***Index Terms*—** Acoustic Scene Classification, MobileNet, Multiple Classifiers, Knowledge Distillation, Quantification

1. INTRODUCTION

Task1a[2] in DCASE2021 challenge aims to classify a test audio recording into one of ten known acoustic scene classes. The development dataset of DCASE2021 task1a is exactly the same as that of dcase2020. But the task is more challenging. Not only we need to classify the audio with mismatched device, but also the model size of our system is limited under 128 kilobytes. That is to say, the system non-zero parameters is smaller than 32768 when using float32 data type, and 65536 when using float16 data type. It's extremely challenging to achieve the last year's state-of-the-art result, which is submitted with more than millions or billions of parameter.

This paper describes our four submission for Task1a of DCASE2021. We build a small-size mobilenet baseline with one frequency branch and two classifiers based on the tencent work[1], and than quantify the model to 16-bit float-point after training. By quantifying, the model size can be compressed to half of the original size, which smaller than 128KB.

2. PROPOSED SYSTEM

2.1. Acoustic Feature Extraction

The audio recording of task1a is provided in single channle with 44.1Hz and 24-bit resolution. We use the librosa library to generate the log-mel spectrogram. Similar to our previous work[3], we first get the STFT spectrogram by using window size of 2048 point, a hop-length of 1024 and Hann window function. Then we use 128 bins mel filter bank to get the log-mel spectrogram of size(128, 431). Additionally, deltas and delta-deltas were calculated from the log-mel spectrogram and stacked into the channel axis, to form the final feature with shape(128, 423, 3).

2.2. The Baseline Model

The task1b of DCASE2020 aims to build the system with low complexity. So we follow the idea of the top2's[1] work. Their mobnet is two frequency branches, which aim to learn the low frequencies and the high frequencies separately. But the model size is too large for this task1a. To build our baseline model, we merge the two branches into one branch and reduce the number of filter in convolution layers.

In order to improve the accuracy, we added a auxiliary three-class classifier which tried to classify the audio recording into three main classes,including in-door, out-door and transportation. Then the baseline model were jointly optimized by the two classifiers. They both use the cross-entropy loss with the loss weights(0.9, 0.1). So the Total Loss = 0.9 × ten-CE Loss + 0.1 × three-CE Loss.

2.3. Knowledge Distillation

Knowledge distillation[4] is widely used in model compression and transfer learning, which can refine the knowledge of multiple models into a single model. In distilling, there are a teacher model which is the exporter of knowledge with no restrictions on model architecture, parameters and integration, and a student model which is the receiver of knowledge with small parameters and relatively simple model structure. Knowledge distillation usually has two steps. One step is training the teacher model. Two step is using the teacher model result as soft-label to distilling teacher model's knowledge into student model. The structure of the knowledge distillation is shown in Fig1.

The performance of teacher model has a great influence on the student model. We use a 18-layer pre-activation ResNet[5] as our teacher baseline with ten-class classifier and three-class classifier . And we total trained two teacher models.

Teacher Model 1: we trained the teacher baseline using the data augmentation methods mentioned in[1], including mixup, random cropping, spectrum augmentation, spectrum correction, pitch shift, speed change, add random noise and mix audios.

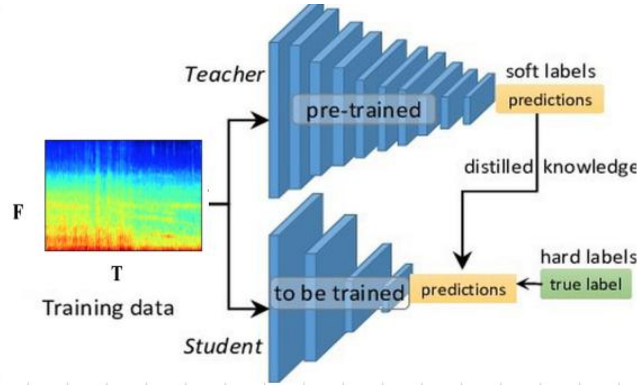

Figure 1: Knowledge distillation structure.

Teacher Model 2:we train the teacher baseline with attention mechanism mentioned in our previous work [3]. And build the ensemble model by averaging the result of 254 and 510 epochs

2.4. Data Augmentation

Data augmentation is a efficient way to avoid overfitting and enhance the model's generalization in deep neural networks. We use mixup [6] and random crop for data augmentation. The mixup alpha equal to 0.4. The random crop length along the time axis is set to 400. Then the final feature sent to network is cropped into 128 × 400 × 3.

2.5. Training Setup

We trained our model using Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9. The initial learning rate was set to 0.1 and controlled by a cosine annealing schedule[7] and restarts with initial learning rate at 2, 6, 14, 30, 126, 254 and 510 epochs. The minimum learning rate was set to 1e-5.

2.6. Model Ensemble And Quantification

Due to the limitation of model size, we didn't use the results of different epochs to make ensemble prediction. Instead, we use the average or weighted average of the model weights in different epochs. After training, we use the tensorflow lite[8] tool to quantify the model from 32-bit float-point to 16-bit float-point, which compresses the model size to half of it's original size with a minor performance drop. The Final model size was compressed to 116KB.

3. RESULTS

The evaluation results of our systems training only on the train split are shown in Table 1. System 1 was distillated from the teacher model 1 using all labels from teacher model and the results were only from the ten-class classifier. System 2-3 distillated from the teacher model 2 as the way of system1, but the results of system2 were only from the ten-class classifier and the results of system3 were the combination of the ten-class classifier and the three-class classifier . System 4 was distilled from the teacher model 2 using the ten-class labels from the teacher model and the three-class labels from the ground truth.

Table 1: Class-wise accuracy for the development dataset

Class	System ID			
	1	2	3	4
airport	51.4%	56.8%	56.8%	50.7%
bus	74.7%	71.7%	73.4%	69.4%
metro	51.5%	57.2%	57.9%	50.5%
metro_station	53.5%	57.9%	57.6%	64.0%
park	71.7%	75.8%	76.1%	68.7%
public square	37.7%	39.4%	38.7%	33.3%
shopping mall	48.8%	52.2%	55.2%	60.9%
street pedestrian	42.8%	41.8%	39.1%	37.4%
street traffic	84.5%	84.2%	83.2%	81.1%
tram	59.1%	59.1%	60.8%	62.2%
Avg accuracy	57.6%	59.6%	59.9%	57.8%
# of parameters	116KB	116KB	116KB	116KB

The systems we submit were trained on the total development.

4. CONCLUSIONS

In this technocal report, we submit four acoustic scene classification systems with small parameters. We use the knowledge distillation to improve the performance of the model, and quantification to reduce the model size. We achieved an accuracy of 59.9% with a model size smaller than 128KB.

5. REFERENCES

[1] Hu H , Yang C , Xia X , et al. "Device-Robust Acoustic Scene Classification Based on Two-Stage Categorization and Data Augmentation." DCASE2020 Challenge, Tech. Rep., June 2020.

[2] http://dcase.community/challenge2021/.

[3] Jie Liu. "Acoustic Scene Classification with Residual Networks and Attention Mechanism." DCASE2020 Challenge, Tech. Rep., June 2020.

[4] Hinton, G. , O. Vinyals , and J. Dean . "Distilling the Knowledge in a Neural Network." *Computer Science* 14.7(2015):38-39.

[5] Mark D. McDonnell and Wei Gao, "Acoustic Scene Classification Using Deep Residual Networks with Late Fusion of Separated High and Low Frequency Paths," DCASE2019 Challenge, Tech. Rep., June 2019.

[6] H. Zhang, M. Cisse, Y. N. Dauphin, and D. LopezPaz, "mixup: Beyond Empirical Risk Minimization," in arXiv:1710.09412, 2017.

[7] Loshchilov, I. , and F. Hutter . "SGDR: Stochastic Gradient Descent with Warm Restarts." *ICLR 2017 (5th International Conference on Learning Representations)* 2016.

[8] M. Abadi, A. Agarwal, P. Barham, et al. "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/