MINI-SEGNET FOR LOW-COMPLEXITY ACOUSTIC SCENE CLASSIFICATION

Technical Report

Yunfei Shao[1], Xuan Zhang[2], Gege Bing[1], Kemeng Zhao[1], Junjie Xu[2],Yong Ma[2], Wei-Qiang Zhang[1]
1. Beijing National Research Center for Information Science and Technology
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
2. School of Lingustic Sciences and Arts, Jiangsu Normal University, Xuzhou 221116, China
shaoyf@tsinghua.edu.cn, zx14xinkestudent@126.com, bgg19@mails.tsinghua.edu.cn, 2120339278@qq.com, 2847419445@qq.com,
may@jsnu.edu.cn, wqzhang@tsinghua.edu.cn

ABSTRACT

This report details the architecture we used to address task 1 of the DCASE2022 challenge. The goal of the task is to design an audio scene classification system for device-imbalanced datasets under the constraints of model complexity. Our architecture is based on SegNet, adding an instance normalization layer to normalize the activations of the previous layer at each step. Log-mel spectrograms, delta features, and delta-delta features are extracted to train the acoustic scene classification model. A total of 6 data augmentations are applied as follows: mixup, time and frequency domain masking, image augmentation, auto level, pix2pix, and random crop. We apply three model compression schemes: pruning, quantization, and knowledge distillation to reduce model complexity. The proposed system achieves higher classification accuracies and lower log loss than the baseline system. After model compression, our model achieves an average accuracy of 54.11% within the 127.2 K parameters size, 8-bit quantization, and MMACs less than 30 M.

Index Terms— Acoustic scene classification, SegNet, data augmentation, model compression

1. INTRODUCTION

Acoustic Scene Classification (ASC) [1] is a task of classifying the acoustic scene presented by given audios. It is a multi-class classification task recognizing the recorded environment sounds specific acoustic scenes that characterize either the location or situation such as park, metro station, tram, etc. Recently, developing signal processing methods to automatically extract audio information has great potential in many application fields, such as searching multimedia based on audio content, manufacturing context aware mobile devices, and intelligent monitoring systems. Further, we aim for the task1 which becomes more challenging in comparison with it was last year on account of stricter restrictions on model size and shorter audio data.

As one of the substantial tasks, acoustic scene classification has been extensively practiced in every challenge. DCASE 2018 and 2019 proposed the mismatch in different recording devices A, B, C and D. Then in 2020 and 2021, the task of acoustic scene classification was divided into 2 subtasks. Among them, the subtask A worked on the dataset collected with mismatched recording devices in 2020 and added requirements for model complexity in 2021. In 2022, the task 1[2] has no subtasks, but it has stricter restrictions on the complexity of the model than in previous years, such as the model’ s number of parameters and the multiply-accumulate operations count. Especially, the audio files have a length of 1 second instead of 10 second therefore 10 times more files than in the 2020 version.

Over these years, the major network structure that has been adopted is residual network based on convolution neural network (CNN) [3-5]. Nevertheless, the performance of ResNet decreases since the audio length is shortened to 1s this year. For task 1, we adopt a mixed semantic segmentation network SegNet, which is improved on our mini-SegNet in 2020 [6-7]. The mini-SegNet systems consist of two important stages. Firstly, mono audio signals are converted to time-frequency representations, scaled by spectrum correction, and zero mean and unit variance normalization. Secondly, the log-mel feature is fed to Mini-SegNet model for feature learning. Contrast to mini-SegNet system, we have made some important adjustments and improvements this year. At the stage of data processing, auto level and pix2pix methods are used for data augmentation with the log-mel technique applied to extract features. Furthermore, the dimension of features is doubled by second-order difference, as a basis for random tailoring during subsequent training. For the purpose of meeting the demands of model complexity, knowledge distillation is employed to lessen the system parameters.

The rest of the paper is organized as follows. Section 2 presents our systems, including data processing and proposed model, data augmentation, and model compression. Section 3 provides experiments and the performance of the proposed approach. Finally, conclusion is provided in Section 4.

2. THE SYSTEM

2.1. Data Preprocessing

The one-second audio segments are formatted with a single channel, 44.1kHz sampling rate, and 24-bit resolution per sample, and in the audio preprocessing stage, we present the spectrum of audios in the log-mel domains. We also used second-order differencing for the feature maps.

This work is supported by a grant from the Guoqiang Institute, Tsinghua University. The corresponding author is Wei-Qiang Zhang.

In our work, we transformed audio data into a power spectrogram by skipping every 1024 samples with 2048 length Hanning window. A spectrum of 44 frames was yields from 1 seconds audio file, and each spectrum was compressed into 256 bins of mel frequency scale. Additionally, deltas and delta-deltas were calculated from the log Mel spectrogram and stacked into the channel axis. The number of frames of the input feature was cropped by the length of the deltadelta channel so that the final shape becomes [256 × 36 × 3]. We replicated the features in the time dimension as double the original to improve the accuracy. During training, random clipping is applied to the temporal dimension of the training set features. In our experiments, the input data with the size of [256 × 72 × 3] is cropped into [256 × 64 × 3] input feature map and a test set of dimension [256×36 ×3].

2.2 Proposed Model

The complete network architecture is presented in Figure 1, which is a teacher model in knowledge distillation. It contains 474K parameters in total.

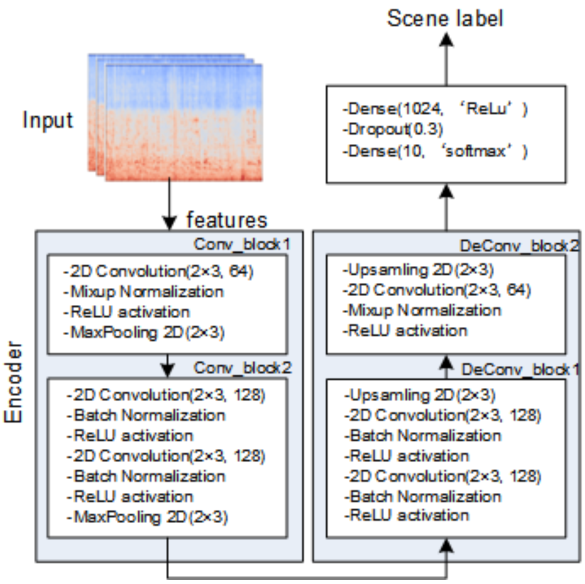

Figure 1: Mini-SegNet architecture

We design a mini-segnet model for low-complexity acoustic scene classification. It is mainly composed of encoder and decoder modules. In both modules, we choose ReLU as the nonlinear activation function.

The encoder module consists of two Conv blocks. The first Conv block contains a convolution layer with a kernel of size 2 ×3, followed by a mixup normalization layer, a non-linear activation layer, and a max-pooling layer. The second Conv block contains two convolution layers with kernels of size 2×3, each followed by a batch normalization layer, a non-linear activation layer, and a max-pooling layer. The output of the encoder is then taken as the input of the decoder module.

The decoder module consists of two DeCov blocks as counterparts to the two blocks in the encoder module. In the first DeConv block, up-sampling is performed first, followed by a convolution layer, a batch normalization layer, and ReLU activation, and then we put a repetition of these three layers above for better performance. The second DeConv block contains four up-sampling layers:2D convolution, mixup normalization, and ReLU activation.

To avoid overfitting, we put a dropout layer after a dense layer. two dense layers are utilized to output final predictions.

2.3 Data Augmentation

To improve the generalization of the model and to prevent overfitting, various data augmentation methods were used.

2.3.1. Mixup, Noise Reduction, Image Augmentation

Mixup [8] is a simple and effective method of data augmentation, which significantly improves results in several areas. Mixup produces new sample-labeled data by summing two sample-labeled data pairs proportionally. In the present work, we use mixup with alpha equal to 0.3. Mixup is performed at a mini-batch level: two data batches, along with corresponding labels, are randomly mixed in each training step. Mixup creates a new training sample by mixing a pair of two training samples. Create a new training sample$(X, y)$from the data and label pair (X_1, y_1) (X_2, y_2) by the following equation:

$$\begin{aligned} X &= \lambda X_1 + (1-\lambda) X_2 \\ y &= \lambda y_1 + (1-\lambda) y_2 \end{aligned} \tag{1}$$

where λ ∈ [0,1] is acquired by sampling from the beta distribution $B \in (\alpha, \alpha)$, and α is a hyper parameters. Besides the data X_1 and X_2, it is characteristic to mix the labels y_1 and y_2.

We use a mask-based method for noise reduction in the time-frequency domain. After extracting the frequency domain features from the audio, we calculate a frequency domain mask by computing the frequency domain features of the clean audio and the corresponding noise-added frequencies, and then train with the noise-added data, using the mask as a label. In this work, we use two frequency masks and a temporal mask with mask parameters of 20 and 3.

We also transfer the image augmentation approach to audio data, using ImageDataGenerator to audio data augmentation to make it more suitable for processing.

2.3.2. Auto Level，Pix2pix

Auto Level [9] is a common data augmentation method in image processing, which we have transferred to audio data augmentation. The comparison of the language spectrum map of the real device before and after using auto level and the language spectrum map of the virtual device is shown in Fig 2.

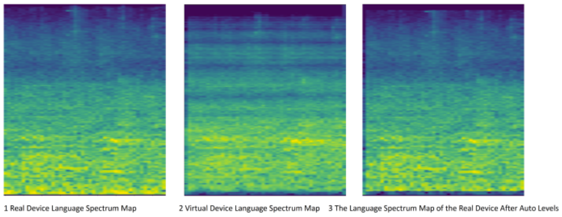
1 Real Device Language Spectrum Map
2 Virtual Device Language Spectrum Map
3 The Language Spectrum Map of the Real Device After Auto Levels

Figure 2: Auto Level processing language spectrum map.

Table 1: Teacher model accuracy on the development dataset**.** For the teacher model (before knowledge distillation), we explore the effects of different augmentations. The results are based on official training and test dataset.

Method	**#Params(total)**	**Average**	**Dev.A**	**Dev.B**	**Dev.C**	**Ave.S1-S3**	**Ave.S4-S6**
Mini-SegNet	474k	58.14%	68.10%	62.95%	66.66%	56.00%	51.71%
Mini-SegNet + Auto Level		59.06%	68.24%	63.86%	66.93%	58.45%	52.35%
Mini-SegNet +Auto Level +pix2pix		60.58%	69.42%	62.80%	67.23%	58.40%	55.06%

Pix2pix [10] is a kind of generative adversarial network. We can use this model to Neural Style. Pix2pix will learn two sets of corresponding images and generate new images.In our work, we use unet networks as generators of pix2pix networks.

In this experiment, we use the spectrograms of recorded audio from real devices and the spectrograms of generated audio from the corresponding virtual devices as the training set for pix2pix. By doing so, we can generate new data, improve the generalization ability of the model, and enhance the accuracy of identifying virtual device-generated data. The result of data generation is shown in the figure 3.

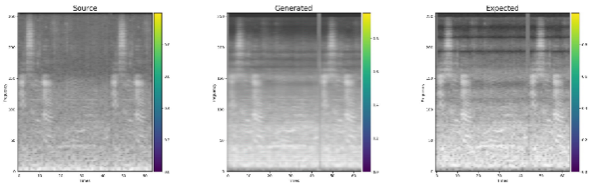

Figure 3: Pix2pix generated feature maps.

2.4 Model Compression

Three model compression methods to reduce our model complexity: network pruning, quantization and knowledge distillation.

Pruning can effectively produce much smaller but faster and more memory-efficient computational models with minimal loss of accuracy. Model quantization is a model compression technique that converts floating-point storage (operations) to integer storage (operations). Knowledge distillation[12] is an effective method based on the "teacher-student network idea", where a simplified model(denoted as student model) can have a relatively good performance under the guidance of a more complex model(denoted as teacher model).

After model compression, our model meets the requirements of MMACs less than 30M and the number of parameters less than 128k.

3. EXPERIMENT

3.1 Experiment setup

All trainings were done on GPU, with a batch size of 128, with the cross-entropy loss function. At the same time, we will use a warm restart[11] learning rate schedule, its maximum value of 0.1 after 11.0, 31.0, 71.0, 151.0, and 311.0 epochs, and then decays according to a cosine pattern to $1 \times 10-5$. In our work, each network has trained for 310 epochs. During the training stage, we use different data augmentation methods for the dataset for Mini-SegNet, such as Mixup with $\alpha = 0.3$, ImageDataGenerator with width shift range = 0.6 and height shift range = 3, and Specaugment with a temporal mask and two frequency masks with mask parameters of 3 and 20, respectively. Experiments show that this method can improve the accuracy of acoustic scene classification.

3.2 Results and discussion

The experimental results and details of submissions can be confirmed in Table 1~3.

According to the results in table 1, auto level provides a relatively higher accuracy for audio from each device, while pix2pix does more for improving the performance of virtual devices(S4-S6). On average, a combination of different augmentation methodology generates a better performance of the model training.

Table 2. Student model accuracy. Three architectures are trained with the Mini-SegNet as teacher model. In the Full Dev.data column, (O) represents using the whole development dataset for training, and (X) represents only official training dataset.

Model	**Full Dev. data**	**#Params(total)**	**Avg. Acc**
1	teacher (X) student (X)	120.21k	53.09%
2	teacher (O) student (X)	127.16k	54.12%
3	teacher (O) student (X)	126.07k	52.83%

In table 2, three architectures are trained with the Mini-SegNet as teacher model. The total parameter size is reduced below 130k after knowledge distillation, achieving an average accuracy of 53%±1%.

In the final submission shown table 3, we ensemble our various experimental schemes to further improve the system generalization ability. Model ensemble is successful in boosting

Table 3. After ensembled all the kinds of subsystems, accuracy of different devices on the development dataset (for device A, B, C, average B&C, average S1-S3, and average S4-S6).

Ensemble	Average ACC	Dev.A	Dev.B	Dev.C	Ave.S1-S3	Ave.S4-S6
task1_1	57.42%	68.27%	62.04%	65.81%	55.23%	51.68%
task1_2	56.68%	67.48%	60.94%	66.02%	54.94%	50.61%

the system's performance according to previous experiments. We ensemble our models using linear combination as follows:

$$y_{\text{ensemble}} = \sum_{n=1}^{N} w_n y_n + b$$

where N is the number of subsystems, y_n is the output score of each subsystem, w_n is the weight coefficient for each subsystem, and b is the bias. The detailed accuracy after fusion is shown in table 3. We submitted two prediction results using different weights. task1_1 ensembled three student models and achieved an average accuracy of 57.42% on the development dataset, while task1_2 ensembled two different student models and achieved an average accuracy of 56.68% on the development dataset.

4. CONCLUSIONS

In this report, we present the methods and techniques we use in the task 1 of the DCASE2022 challenge. We extract log-mel spectrograms, delta and delta-delta features and apply various techniques for normalization and augmentation. To design an efficient acoustic scene classification model, we adopt Mini-SegNet with an instance normalization layer after each step. We compress the model further by utilizing three model compression schemes: pruning, quantization, and knowledge distillation. Our system achieves 54.11% accuracy within the 127.2 K parameters .

5. REFERENCES

[1] https://dcase.community/

[2] https://dcase.community/challenge2022/task-low-complexity-acoustic-scene-classification

[3] Kim B, Yang S, Kim J, et al. QTI submission to DCASE 2021: Residualnormalization for device-imbalanced acoustic scene classification withefficient design[R]. DCASE2021 Challenge, 2021

[4] Hee-Soo H, Jee-weon J, Hye-jin S, et al. Clova submission for the DCASE2021 challenge: Acoustic scene classification using light architecturesand device augmentation[R]. DCASE2021 Challenge, 2021

[5] Byttebier L, Desplanques B, Thienpondt J, et al. Small-footprint acousticscene classification through 8-bit quantization-aware training andpruning of ResNet models[R]. DCASE2021 Challenge, 2021.

[6] Badrinarayanan, V., Kendall, A., Cipolla, R., "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence , 2015.

[7] Xinxin Ma, Yunfei Shao, Yong Ma, Wei-Qiang Zhang. "Deep semantic encoder-decoder network for acoustic scene classification with multiple devices," In: Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Auckland, New Zealand, pp. 365–370 , 2020.

[8] H. Zahng, M. Cisse, Y. N. Dauphin, and D. Loped -paz, "mixup: beyond empirical risk minimization," arxiv preprint arxiv:1710.09412, 2017

[9] Hongbo Feng, Ping Li, Bo Li. Color Image Enhancement Algorithm Based on Multi-scale Retinex and Automatic Color Method[J]. Radio Engineering, 2019, 49(10): 910-914.

[10] Isola P , Zhu J Y , Zhou T , et al. Image-to-Image Translation with Conditional Adversarial Networks[C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2016.

[11] I. Loshchilov and F. Hutter, "SGDR: stochastic gradient descent with restarts," CoRR, vol. abs/1608.03983, 2016. [online]. Available: http://arxiv.org/abs/1609.03983

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