BIT SUBMISSION FOR DCASE 2020 CHALLENGE TASK1

Technical Report

Yuxiang Wang†, Shuang Liang†,Qingran Zhan†, Xiang Xie†¶,

Information and Electronics Institute, Beijing Institute of Technology, Beijing, China‡
Shenzhen Research Institute, Beijing Institute of Technology, Shenzhen, China¶

ABSTRACT

DCASE2021 challenge task 1 contains two sub-tasks: (i) Low-Complexity Acoustic Scene Classification with Multiple Devices and (ii) Audio-Visual Scene Classification. In our submission systems, different methods are used for different tasks. For task 1a, we explore the fsFCNN (frequency sub-sampling controlled fully convolution) with two-stage training. In order to reduce the model size, the knowledge distillation approach is used. For task 1b, different models are used for different modals. For audio classification, the same fsFCNN structures with two-stage training are applied. For video classification , the TimeSformer model is used.

Experimental results show that our final model obtain accuracy of 64.6% with 128kb model size. On task 1b development set, the audio modal achieve 80.6% accuracy and 92% for video modal.

Index Terms— acoustic scene classification, log-mel spectrograms, Convolutional Neural Network, Transformer

1. INTRODUCTION

DCASE2021 challenge task1 contains two tasks which are focused on solving different problems. The goal of the task1a is to classify acoustic scenes in situation of multiple recording devices with limited model size. For task 1b, it is concerned with classification using audio and video modalities.

In our submitted system, we use the fscnn with two-stage training and for task 1b, the same fscnn is used for audio classification and for video classification, the TimeSformer [1] is adapted on the ASC task.

The details of the experiments setup will be described in the following parts.

2. MODEL ARCHITECTURES

2.1. Model for task 1a

ASC algorithms mostly use convolutional neural network (CNN) based network architectures since they usually provide a summarizing classification of longer acoustic scene excerpts [2]. We follow the idea of [3] and use the model of FCNN and fsFCNN as the baseline.

In order to realize low-complexity solutions, knowledge distillation is used. We use FCNN and fsFCNN fusion network as the teacher net and a simple five-layer CNN network as student net.

2.2. Models for task 1b

For the Multimodal task, different models for different modal are considered.

For the audio, we follow the idea of [3] and use the two-stage classifiers. Two different CNN based models Resnet and fsFCNN are used in the two-stage classifier.

For the video, the TimeSformer is used. TimeSformer is a convolution-free approach to video classification built exclusively on self-attention over space and time [1]. The model adapts the standard Transformer architecture to video by enabling spatiotemporal feature learning directly from a sequence of frame-level patches.

3. EXPERIMENTS SETUP

3.1. Data preparation

In task 1a, The dataset contains data from 10 cities and 9 devices: 3 real devices (A, B, C) and 6 simulated devices (S1-S6). Data from devices B, C, and S1-S6 consists of randomly selected segments from the simultaneous recordings, therefore all overlap with the data from device A, but not necessarily with each other. The total amount of audioset is 64 hours [4]. Meanwhile, the dataset is provided with a training/test split in which 70% of the data for each device is included for training, 30% for testing. The audio-video dataset is task 1b is TAU Audio-Visual Urban Scenes 2021. The dataset contains synchronized audio and video recordings from 12 European cities in 10 different scenes. The provided audio is recorded using a Soundman OKM II Klassik/studio A3, electret binaural microphone and a Zoom F8 audio recorder using 48kHz sampling rate and 24-bit resolution. The provided video is recorded using a GoPro Hero5 Session. Faces and licence plates in the video were blurred during the data postprocessing stage [5].

For the audio files in both tasks, the Log-mel filter bank features were used in our experiments. The input audio waveform is analyzed with a 2048 SFFT points, a window size of 2048 samples, and a frameshift of 1024 samples. The librosa [6] library is used to extract the features.

A key element of our submission is how to do data augmentation strategies. For all the audio files in task 1a and 1b, the following data augmentation methods are used [3]: Mixup, Random cropping, spectrum augmentation, spectrum correction, pitch shift, speed change, random noise and mix audios.

3.2. Model trianing

For the audio modal in both tasks, Stochastic gradient descent (SGD) with a cosine-decay-restart learning rate scheduler is used to train our models.

For the video modal in task1b, we use ViT architecture to pretrain on ImageNet for image preprocessing.

4. CONCLUSION

In this technical report, we proposed different methods for task1a and task1b. For the audio modal, we use the CNN based models and multiple enhancement methods to improve the preformance of the system. For video classification , the TimeSformer model is used.

For task1a, our final model obtain accuracy of 64.6% with 128kb model size which is 17% over than the baseline system. On task 1b development set, the audio modal achieve 80.6% accuracy and 92% for video modal.

5. REFERENCES

[1] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" *CoRR*, vol. abs/2102.05095, 2021. [Online]. Available: https://arxiv.org/abs/2102.05095

[2] J. Abeßer, "A review of deep learning based methods for acoustic scene classification," *Applied Sciences*, vol. 10, no. 6, 2020.

[3] H. Hu, C.-H. H. Yang, X. Xia, X. Bai, X. Tang, Y. Wang, S. Niu, L. Chai, J. Li, H. Zhu, F. Bao, Y. Zhao, S. M. Siniscalchi, Y. Wang, J. Du, and C.-H. Lee, "Device-robust acoustic scene classification based on two-stage categorization and data augmentation," 2020.

[4] 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

[5] S. Wang, A. Mesaros, T. Heittola, and T. Virtanen, "A curated dataset of urban scenes for audio-visual scene analysis," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, accepted. [Online]. Available: https://arxiv.org/abs/2011.00030

[6] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," 2015.