DCASE 2024 CHALLENGE TASK10 TECHNICAL REPORT

Technical Report

Zhilong Jiang, Xichang Cai[*],Ziyi Liu, Menglong Wu*

North China University of Technology, Beijing, China
Caixc_ip@126.com

ABSTRACT

This technical report describes our approach to Challenge 10 of DCASE 2023: acoustic based traffic monitoring. In our work, we use Mel spectrogram and Vgg11 algorithm to extract category features of vehicles in sound, while using GCC-PATH algorithm and CNN algorithm to extract directional features of vehicles in sound. Meanwhile, we also optimize the experimental results by continuously adjusting the parameters in the algorithm.

***Index Terms*—** Mel, Vgg11, CNN, GCC-PATH

1. INTRODUCTION

The goal of Task 10 of the DCASE 2023 challenge is to use a given traffic audio clip to distinguish the categories of vehicles and their rotation directions within each minute of sound. In this work, [2,3] we first enter the Harmonoise model with the given car engine sound, and the generated signal is used as input for Pyroadacoustics. Finally, we obtain simulated sound data to generate a pre trained model. Meanwhile, use real-world sound data to fine tune the model. In pre training and fine-tuning, Mel spectrogram and Vgg11 algorithm are used to extract category features of vehicles in sound, while GCC-PATH algorithm and CNN algorithm are used to extract directional features of vehicles in sound. The remaining parts of this report are organized as follows. Section 2 provides a detailed introduction to our method and the materials submitted. Section 3 shows the experimental results.

2. OUR APPROACHES AND SUBMISSIONS

2.1 Vgg11

The VGG11 algorithm is a classic deep convolutional neural network (CNN) architecture, which consists of five convolutional blocks, each consisting of multiple convolutional layers and a pooling layer. The first convolutional block has a convolutional layer that outputs 64 channels; The next two convolutional blocks each have two convolutional layers, with output channels of 128 and 256, respectively; The last two convolutional blocks each have three convolutional layers, with an output channel count of 512. [4] Meanwhile, the VGG11 network has a large number of parameters, totaling approximately 133 million parameters. These parameters are mainly distributed in convolutional layers and fully connected layers. Due to the use of smaller convolutional kernels (3x3) and deeper network structures, VGG11 has a relatively large number of parameters but also stronger feature representation ability.

2.2 GCC-PATH

The GCC-PHAT algorithm is based on the principle of generalized cross correlation, which estimates the delay between signals by calculating the cross-correlation function between two signals. This algorithm is particularly suitable for time delay estimation in noisy environments, as it suppresses noise and reverberation interference through phase transformation weighting, thereby improving the accuracy of time delay estimation.[5] At the same time, it also has advantages such as strong noise resistance, high computational efficiency, and wide applicability, and has broad application prospects in the fields of sound source localization and signal processing.

3. EXPERIMENTAL RESULTS

During the training process, we used 100 epochs with a batch size of 8 and an initial learning rate of 10 ^ -3. We used the Adam optimizer and used the same loss function as the baseline. The evaluation results of the DCASE2024 Task10 dataset are shown in Table 1.

Loc1	car_left	car_right	cv_left	cv_right
Kendall's Tau Corr	0.424	0.439	0.144	0.107
RMSE	2.614	2.955	0.982	0.890

Loc2	car_left	car_right	cv_left	cv_right
Kendall's Tau Corr	0.497	0.267	0.041	-0.069
RMSE	3.289	3.121	0.846	0.630

Loc3	car_left	car_right	cv_left	cv_right
Kendall's Tau Corr	0.538	0.566	0.155	0.292
RMSE	1.750	1.301	0.334	0.212

Loc4	car_left	car_right	cv_left	cv_right
Kendall's Tau Corr	0.073	-0.164	-0.141	0.411
RMSE	2.138	1.743	0.655	0.463

[*] These authors contributed equally to this work.

Loc5	car_left	car_right	cv_left	cv_right
Kendall's Tau Corr	0.348	0.373	0.053	0.283
RMSE	0.921	0.698	0.383	0.282

Loc6	car_left	car_right	cv_left	cv_right
Kendall's Tau Corr	0.783	0.677	0.740	0.620
RMSE	1.777	1.921	0.531	0.552

Table 1: Results on DCASE2024 Task10 dataset

4. CONCLUSIONS

This article introduces the system we submitted for task 10 of DCASE2024. The system utilizes Mel spectrogram and Vgg11 algorithm to extract category features of vehicles in sound, while using GCC-PATH algorithm and CNN algorithm to extract directional features of vehicles in sound, resulting in better performance than baseline.

5. REFERENCES

[1] http://dcase.community/workshop2024/.

[2] Stefano Damiano and Toon van Waterschoot. Pyroadacoustics: a Road Acoustics Simulator Based on Variable Length Delay Lines. In Proceedings of the 25th International Conference on Digital Audio Effects (DAFx20in22), 216–223. Vienna, Austria, September 2022.

[3] Stefano Damiano, Luca Bondi, Shabnam Ghaffarzadegan, Andre Guntoro, and Toon van Waterschoot. Can synthetic data boost the training of deep acoustic vehicle counting networks? In Proceedings of the 2024 International Conference on Acoustics, Speech and Signal Processing (ICASSP) (accepted). Seoul, South Korea, April 2024.

[4] Simonyan, Karen and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." CoRR abs/1409.1556 (2014): n. pag.

[5] D. Liu, X. Cai, D. Yu, Z. Qiao, H. Dong and M. Wu, "Sound Source Localization Methods Based on Lagrange-Galerkin Spherical Grid," 2021 IEEE International Conference on Electrical Engineering and Mechatronics Technology (ICEEMT), Qingdao, China, 2021.