Fine-Tuning Using Grid Search & Gradient Visualization

Technical Report

Bowei Hou *

Waseda University
Graduate School of IPS, 2-7 Wakamatsu,
Kitakyushu, Fukuoka 808-0135, Japan
houbowei@fuji.waseda.jp

Kacper Radzikowski *

Waseda University
Graduate School of IPS, 2-7 Wakamatsu,
Kitakyushu, Fukuoka 808-0135, Japan
kradziko@fuji.waseda.jp

Ahmed Farid

Waseda University
Graduate School of IPS, 2-7 Wakamatsu,
Kitakyushu, Fukuoka 808-0135, Japan
ahmed.farid@asagi.waseda.jp

ABSTRACT

In this technical report, we briefly describe the models used in the task 4 challenge of DCASE2020. We utilized previously available models and fine-tuned them using the grid search algorithm and gradient visualization. This is the first attempt by our team to enter a competition on sound source manipulation.

Index Terms— DCASE2020, fine-tuning, data augmentation

1. INTRODUCTION

Ubiquitous computing has been very effective in improving our daily lives. This is observable in aspects such as home automation, healthcare, manufacturing, and so on. One of the interesting interfaces for ubiquitous computing is sound. For example, a computer can understand requests and even engage in conversations by means of processing human voice. However, sound scenes tend to be complex; background noise and overlapping sound events can hinder the detection and interpretation of target sounds. The interpretation of targets within complex scenes can be achieved by sound source separation and detection tasks. For single-channel sounds, state-of-the-art approaches utilize DNNs for both the separation and detection tasks. DNNs can be designed to handle each task separately, or they can be designed and trained to handle both tasks as end-to-end. Because the focus of DCASE2020 task 4 is the utilization of single-channel sounds, we aim to build upon existing works based on DNNs.

2. MODEL DESCRIPTION

In this section, we describe the model and efforts made into producing the results. For this submission, we opted for two separate models for sound source separation and sound event detection. Both models were trained separately and were not placed in an end-to-end configuration.

2.1. SOUND SOURCE SEPARATION

The solution for the sound source separation model was based on the provided baseline for the task [1]. The model has been trained using the FUSS reverbant and dry datasets [2]. The fine tuning of the model was performed using the Grid Search algorithm, which was implemented using Scikit Learn library [3] in conjunction with the Keras deep learning library.

2.2. SOUND EVENT DETECTION

The model we utilized is based on the provided baseline model for sound event detection[4][5]. The training was done on the DESED dataset and soundbank training from DESED. We augmented the dataset using synthetic data generated with the Scaper library, similarly to [6]. In order to fine tune the baseline model, we used the Weights and Biases web platform and applied gradient visualization [7].

3. RESULTS

We present our validation results in tables 1 and 2, for sound separation and detection respectively. We directly compare our results to the contest announced baseline results.

Table 1 Sound separation results on the reverberant FUSS dataset

		Val (db)	Eval (db)
Baseline			
Reverbant	Single-source SI-SNR	35.0	37.6
	Multi-source SI-SNRi	13.0	12.5
Dry	Single-source SI-SNR	30.6	31.8
	Multi-source SI-SNRi	10.5	10.2
Ours			
Reverbant	Single-source SI-SNR	35.0	37.6
	Multi-source SI-SNRi	13.1	12.9
Dry	Single-source SI-SNR	30.6	31.8
	Multi-source SI-SNRi	10.6	10.3

* Denotes equal contribution

Table 2 Sound detection results on the DESED dataset

	Baseline	Ours
	Overall	
F-score	34.8	40.84
PSDS	0.610	0.596
	Class-wise F-score	
Alarm-bell ringing	36.1	42.7
Blender	35.2	35.5
Cat	45.1	44.6
Dishes	25.7	24.4
Dog	22.1	29.0
Electric shaver	37.6	31.4
Frying	24.1	23.6
Running water	33.4	28.4
Speech	50.9	52.3
Vacuum cleaner	45.7	42.5

4. CONCLUSION & FUTURE WORK

We presented fine-tuning techniques used on the baseline models. We intend to use the experience gained for this submission to deepen our understanding of the topics and provide stronger contributions in the future.

5. REFERENCES

[1] Ilya Kavalerov, Scott Wisdom, Hakan Erdogan, Brian Patton, Kevin Wilson, Jonathan Le Roux, and John R. Hershey. "*Universal sound separation".* In IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 175–179. October 2019. URL: https://arxiv.org/abs/1905.03330.

[2] Scott Wisdom, Hakan Erdogan, Daniel P. W. Ellis, Romain Serizel, Nicolas Turpault, Eduardo Fonseca, Justin Salamon, Prem Seetharaman, John R. Hershey, "What's All the FUSS About Free Universal Sound Separation Data?", 2020, in preparation.

[3] Scikit Learn, "GridSearchCV". URL: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

[4] Lionel Delphin-Poulat and Cyril Plapous. *Mean teacher with data augmentation for dcase 2019 task 4.* Technical Report, Orange Labs Lannion, France, June 2019.

[5] Tarvainen, A. and Valpola, H., 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In Advances in neural information processing systems (pp. 1195-1204).

[6] Eduardo Fonseca, Jordi Pons, Xavier Favory, Frederic Font Corbera, Dmitry Bogdanov, Andrés Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra. "Freesound Datasets: A Platform for the Creation of Open Audio Datasets." International Society for Music Information Retrieval Conference (ISMIR), pp. 486–493. Suzhou, China, 2017.

[7] Tyler Shimko, "Exploring Gradients", Weights & Biases, 2019. URL: https://www.wandb.com/articles/exploring-gradients